UNIVERSITA' DEGLI STUDI DI CAGLIARI

FACOLTA' DI SCIENZE ECONOMICHE, GIURIDICHE E POLITICHE

Corso di Laurea Magistrale in Economia Manageriale

Il machine learning per la previsione degli acquisti online. Il caso Google Store.

Relatore:

Prof. Luca Frigau

Tesi di Laurea di:

Michael Palmas

Anno Accademico 2019 - 2020

Indice

Introduzione

La velocità di cambiamento che caratterizza il mercato attuale, richiede alle imprese che vogliano mantenere un vantaggio competitivo durevole la capacità di individuare nuove opportunità di crescita e reagire tempestivamente a tali cambiamenti. Questo si traduce nel riuscire a identificare nuovi segmenti di mercato, nuove tendenze e preferenze dei consumatori, e assumere le decisioni necessarie a migliorare e ottimizzare l'efficienza dei processi aziendali. I dati, in questo contesto, assumono un ruolo fondamentale, in quanto fonte di informazioni utili a supportare le decisioni del management. La grande mole di dati e la molteplicità di fonti da cui provengono, rendono la loro elaborazione e analisi particolarmente complessa. Per questo motivo è necessario adottare opportuni strumenti che consentano di risolvere tale problematica.

Questo elaborato si occupa di approfondire uno di questi strumenti, il *machine learning* (apprendimento automatico), in quanto ha riscontrato un forte interesse in molte aree scientifiche, nonché nel marketing, finanza e altre discipline aziendali. Il *machine learning* si basa sull'idea che i sistemi possono imparare dai dati, identificare modelli autonomamente e prendere decisioni.

Partendo da una panoramica sulle diverse metodologie esistenti, si mostra l'applicazione di alcune di queste tecniche ad un caso pratico. In particolare, nel primo capitolo una volta stabilita la differenza tra previsione e inferenza, sono illustrate le differenze principali tra le diverse tecniche applicabili con riferimento ai vantaggi e gli svantaggi che ognuna di queste comporta in termini di prestazioni e interpretazione dei risultati. Inoltre, viene trattato l'aspetto relativo alla valutazione delle performance di un modello di apprendimento automatico. Nel secondo, è stato approfondito l'aspetto della classificazione binaria, uno dei problemi tipici del ML, che consiste nell'attribuzione di una determinata categoria ad un'osservazione. Nel caso della classificazione binaria le categorie possibili sono due e solitamente vengono definite come "classe positiva" e "classe negativa". Nel corso del capitolo ci si focalizza su tre metodologie che consentono di risolvere questo problema: regressione logistica, analisi discriminante lineare e support vector machines. Spesso in un set di dati le categorie di classificazione sono rappresentate da un numero di osservazioni estremamente diverso. Tale sbilanciamento causa importanti problematiche nelle performance di un modello di apprendimento automatico, poiché

tende a focalizzarsi sulla categoria prevalente e a ignorare gli eventi rari, che solitamente costituiscono il concetto di interesse. Per tale ragione, nel terzo capitolo sono state descritte alcune soluzioni per affrontare il problema delle classi sbilanciate e migliorare la classificazione delle osservazioni: SMOTE, ROSE, MWMOTE.

Nel quarto capitolo viene riportata l'analisi esplorativa di un dataset reale, relativo alle sessioni di navigazione all'interno dell'e-commerce Google Store in un determinato lasso temporale. Infine, con l'obiettivo di valutare l'efficacia in termini di prestazioni delle metodologie di classificazione e delle tecniche di bilanciamento citate in precedenza, queste sono state applicate al dataset del Google Store, in cui le categorie acquisto/non acquisto sono particolarmente sbilanciate. Per valutare le prestazioni sono state utilizzate diverse misure di performance. Uno degli indicatori più comuni, l'accuratezza, può condurre a risultati fuorvianti, poiché è influenzato fortemente dalla distribuzione delle categorie. È stato identificato un insieme di altri indicatori di performance, il cui funzionamento è indipendente da tale distribuzione.

1. Il machine learning

In un contesto economico in continuo cambiamento, uno dei fattori determinanti la competitività, ma in primis la sopravvivenza dell'impresa è la sua capacità di individuare tempestivamente nuovi segmenti di mercato, nuove tendenze in termini di preferenze e comportamento dei consumatori e riuscire, in funzione di questi, a adattare i propri processi aziendali. In quest'ottica, il dato rappresenta l'elemento chiave per lo sviluppo del business e il successo dell'impresa, tanto che è stato definito il "nuovo petrolio".

Dai dati si possono trarre importantissime informazioni per quanto riguarda l'andamento aziendale. Si possono valutare i rischi e le opportunità di un investimento in relazione alle risorse disponibili, l'andamento dei profitti in un certo periodo di tempo e analizzare le performance dei dipendenti. I dati sono alla base di ogni decisione operativa e strategia.

L'uso dei dati per guidare il processo decisionale non è una novità in ambito imprenditoriale, la differenza rispetto al passato è la mole di dati disponibili, tanto che si parla sempre più spesso di *big data*.

Tuttavia, questo termine ha un significato che va oltre la quantità, ma fa riferimento anche alla varietà e velocità con la quale i dati vengono generati, immagazzinati ed analizzati. Queste rappresentano ciò che viene indicato come le quattro V dei *big data*, alle quali se ne aggiunge una quinta, il valore:

- Volume: negli ultimi anni la quantità e complessità di dati a disposizione è aumentata notevolmente a causa dell'avvento delle nuove tecnologie, a fianco ai tradizionali canali offline. Se da una parte i dati rappresentano il bene più prezioso per l'azienda, dall'altra il loro volume elevato rende complessa l'estrazione delle informazioni necessarie a supportare il processo decisionale.

- Velocità: indica la velocità e frequenza con la quale i dati vengono generati, costringendo le aziende a dover assumere decisioni in tempi molto più rapidi per cogliere le opportunità offerte dal mercato.

- Varietà: la tipologia di dati a diposizione delle imprese è sempre più eterogenea, infatti ai classici dati provenienti dal sistema informativo interno aziendale si aggiungono quelli provenienti da una molteplicità di fonti esterne tra cui e-mail, immagini, video, social media e altro ancora. La loro natura li rende particolarmente

rilevanti per gli obiettivi di business, ma allo stesso tempo l'elevata complessità di analisi ed elaborazione necessità l'adozione di infrastrutture dedicate.

- Veridicità: si riferisce alla possibilità di estrarre informazioni veritiere dai dati, nonostante la presenza di gravi errori, imprecisioni e incompletezza. Le interazioni degli utenti nei social media, ad esempio, non sempre producono dati veritieri. Si pensi a un like messo per errore o ad uno stesso dispositivo utilizzato da persone diverse. È pertanto necessario utilizzare strumenti adeguati a verificarne l'affidabilità.

- Valore: affinché i dati possano produrre un output informativo di successo, divenendo quindi fonte di valore per le imprese, è necessario che alla fase di raccolta seguano una fase di elaborazione e di analisi dei dati realizzate con metodologie per lo più basate sull'utilizzo di algoritmi e modelli statistici.

Tra gli strumenti che consentono di risolvere le problematiche appena evidenziate, uno dei più innovativi è il *machine learning* (apprendimento automatico), un metodo di analisi dei dati che automatizza la costruzione di modelli analitici. A differenza degli strumenti tradizionali, il *machine learning* utilizza un approccio induttivo al fine di formare una rappresentazione della realtà sulla base dei dati che esamina; rappresentazione che è in capace di ottimizzare e migliorare all'aggiungersi di nuovi dati. Il machine learning si basa su regole statistiche, riuscendo a risolvere una molteplicità di problematiche in diversi ambiti.

Uno studio della multinazionale di consulenza strategica McKynsey ha tracciato i primi 120 casi d'uso del machine learning nell'ambito di 12 diversi settori[1], rappresentati nella figura 1.1. Nell'asse Y sono riportati i volumi di dati disponibili, mentre nell'asse X il potenziale impatto del ML sulle attività interessate. La dimensione della bolla riflette la diversità delle fonti di dati disponibili. Le bolle in alto a destra rappresentano le aree in cui l'utilizzo del ML consentirebbe alle aziende di cogliere nuove opportunità nel mercato, quali ad esempio pubblicità personalizzata, ottimizzazione dei prezzi, ottimizzazione delle strategie di merchandising.

[1] McKinsey Global Institute, (2016), The Age Of Analytics: Competing In A Data-Driven World.

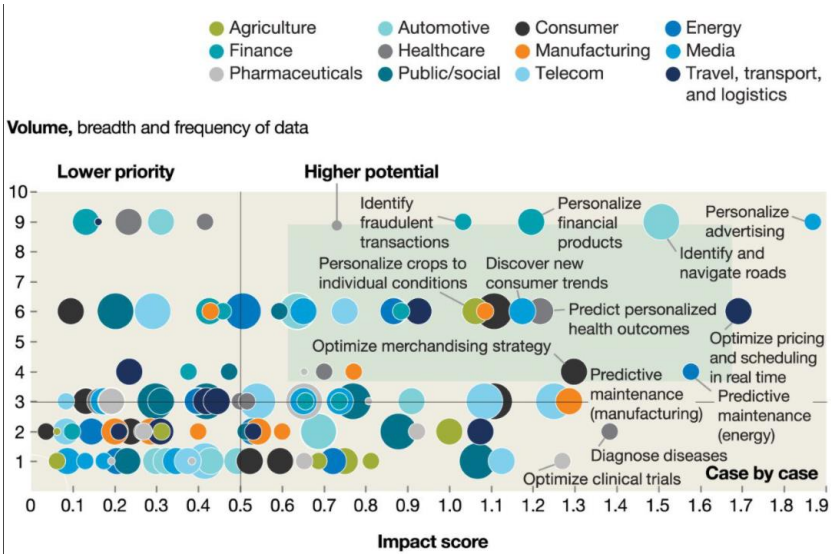


Figura 1.1 L'immagine rappresenta i settori di applicazione del Machine Learning.
Fonte: McKinsey Global Institute, (2016), The Age Of Analytics: Competing In A Data-Driven World.

Nell'ambito dello stesso studio è stato analizzato e misurato l'impatto del ML all'interno dei diversi settori, dei quali si riportano alcuni casi aziendali:

- Paypal, leader nel settore dei pagamenti online, ha costruito dei modelli di ML per identificare i pagamenti fraudolenti in tempo reale. In questo modo sono riusciti a ridurre le perdite degli anni precedenti che ammontavano a circa 10 milioni di dollari al mese.

- Distillery, una società specializzata nel marketing, utilizza i dati di navigazione, quali pagine visitate, click ed acquisti, per elaborare campagne di display advertising, efficienti sia in termini di costo che di prestazioni rispetto a quelle svolte da un singolo individuo.

- Il sistema sanitario Carolinas (CHS) utilizza un modello per assumere le decisioni in merito alla dimissione o meno dei pazienti, riducendo così i casi di successivi ricoveri.

Un ulteriore utilizzo in ambito economico è il caso di un istituto bancario che deve verificare l'affidabilità creditizia di coloro che richiedono un prestito. Il cliente potrà essere considerato affidabile oppure non affidabile, perciò verrà attribuito un risultato sulla base di alcune caratteristiche note. Le informazioni note possono essere rappresentate da variabili numeriche (età, stipendio mensile, patrimonio immobiliare) e da variabili

qualitative (genere, nazionalità, professione). Il risultato è definito come variabile di output o dipendente e solitamente viene indicato con la lettera Y, invece, le caratteristiche note sono definite come variabili di input o indipendenti, indicate dalla lettera X. Il valore che assume una data Y sulla base di *p* differenti input X = (X_1, X_2, X_3, …, X_p), supponendo che esista una relazione tra Y e X, può essere espresso tramite la seguente formula:

$$Y = f(X) + \varepsilon \tag{1.1}$$

Dalla formula emerge che la variabile di output non sia perfettamente definita dagli input, poiché è presente un errore casuale *ε,* indipendente da X e che ha media pari a zero. Inoltre, nonostante le variabili indipendenti forniscano informazioni su Y, la forma della funzione è ignota. Infatti, l'apprendimento automatico si riferisce a una serie di approcci per stimare *f* utilizzando le informazioni note ovvero imparare il meccanismo con il quale da X ricaviamo Y. Una caratteristica comune tra i metodi di apprendimento automatico è l'utilizzo di un insieme di osservazioni, chiamato dati di training. A queste osservazioni si applica il metodo di apprendimento automatico, al fine di insegnare al modello come stimare la funzione ignota *f.* Ci sono due motivi principali per il quale si potrebbe voler stimare *f:* per fare previsioni e per fare inferenza. [2]

1.1 Previsione e inferenza

L'obiettivo della previsione è fornire stime accurate di Y e non si focalizza particolarmente sulla relazione tra input e output. Poiché l'errore è in media pari a zero, tale previsione può essere espressa utilizzando:

$$\hat{Y} = \hat{f}(X) \tag{1.2}$$

in cui $\hat{f}$ rappresenta la stima di *f* e $\hat{Y}$ rappresenta il risultato della previsione di Y.
Generalmente le previsioni sono imperfette a causa delle differenze tra i valori stimati e i valori reali. L'accuratezza delle previsioni è determinata da due quantità: l'errore riducibile e l'errore non riducibile. L'errore riducibile dipende dalla mancata corrispondenza tra $\hat{f}$ e *f.* Tale errore è riducibile, perciò è possibile migliorare l'accuratezza della previsione, attraverso l'utilizzo della tecnica di apprendimento automatico più adeguata. Tuttavia,

[2] James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

anche quando è possibile stimare perfettamente f, la previsione sarà condizionata dall'errore non riducibile. Tale errore deriva dal fatto che X non determina completamente Y. Infatti, Y è anche funzione di ε, indipendente da X, e la sua variabilità influenza l'accuratezza delle previsioni. La componente aleatoria ε potrebbe comprendere alcune variabili utili per prevedere correttamente Y: poiché tali variabili non sono state misurate, non è possibile utilizzarle per la previsione.

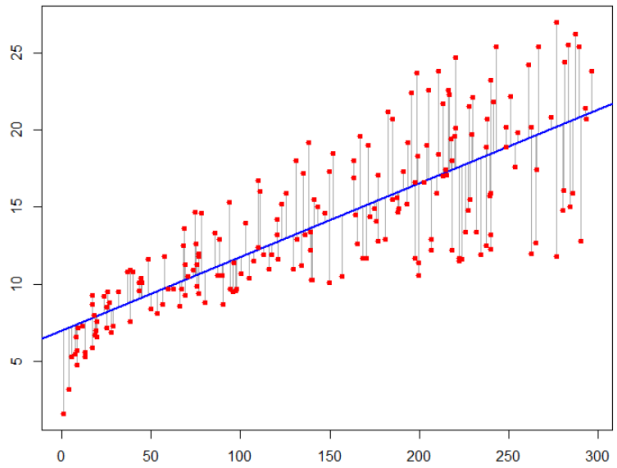


Figura 2.2 La linea blu rappresenta la previsione Ŷ, i punti rossi indicano i valori reali di Y e le linee grigie gli errori di previsione, dati dalla somma dell'errore riducibile e dell'errore non riducibile.

Si consideri una data $\hat{f}$ e un insieme di variabili X, che generano la previsione $\hat{Y} = \hat{f}(X)$. Si assuma che sia $\hat{f}$ che X siano stati fissati.

Si può dimostrare che:

$$E(Y - \hat{Y})^2 = E[f(X) + \varepsilon - \hat{f}(\mathrm{X})]^2 = [f(X) - \hat{f}(X)]^2 + Var(\varepsilon) \qquad (1.3)$$

$[f(X) - \hat{f}(X)]^2$ rappresenta l'errore riducibile, invece, $Var(\varepsilon)$ è la varianza associata all'errore non riducibile. Il risultato $E(Y - \hat{Y})^2$ è la media, o valore atteso, della differenza al quadrato tra il valore previsto ed effettivo di Y.

Per fare inferenza l'obiettivo è capire in quale modo Y cambia in funzione di X_1, X_2, X_3, …, X_n. In questo caso l'interesse è orientato alla possibile relazione tra X e la variabile dipendente Y, perciò è necessario conoscere la forma esatta di $\hat{f}$. Per il raggiungimento di tale obiettivo è rilevante considerare che solo alcune delle caratteristiche note sono sostanzialmente associate a Y, perciò è utile individuare le variabili indipendenti più importanti. Alcuni predittori possono influire positivamente sull'output, cioè un aumento del loro valore può comportare l'aumento del valore di Y, invece, altri possono determinare

variazioni in diminuzione. Ad esempio, qualora Y sia il prezzo di un appartamento, un aumento delle dimensioni comporterebbe un valore più elevato, invece, il tasso di criminalità della zona influirebbe negativamente. A seconda della complessità di f l'effetto generato dalla variazione di un predittore può dipendere anche dal valore di altre variabili. Se si considera l'esempio precedente, la presenza di una piscina in un immobile di grandi dimensioni può generare un aumento di prezzo maggiore rispetto alla variazione che si potrebbe osservare in un appartamento più piccolo. Inoltre, la maggior parte dei metodi per stimare f assumono che abbia una forma lineare. In alcune situazioni, tale ipotesi è ragionevole o auspicabile, ma spesso la relazione tra Y e X è più complessa e un modello lineare non fornisce una rappresentanza adeguata.

In alcune circostanze dagli stessi input si possono risolvere sia problemi di previsione che di inferenza. Ad esempio, in un contesto immobiliare si potrebbe mettere in relazione il prezzo di un appartamento con alcune caratteristiche come la sua dimensione, il tasso di criminalità, la presenza di una scuola, il reddito medio del quartiere e altri input. Qualora l'obiettivo fosse stabilire il valore di un appartamento, date le caratteristiche, si tratterebbe di una previsione. Infatti, l'interesse non è orientato a stabilire la relazione tra le singole variabili e il prezzo, ma solamente a prevedere in maniera accurata il valore dell'immobile sulla base delle caratteristiche definite in precedenza. Invece, se l'indagine fosse rivolta a calcolare l'importo della variazione di prezzo dovuta alla presenza di una scuola, si parlerebbe di un problema di inferenza. A seconda dell'obiettivo è possibile utilizzare metodi diversi per stimare f: lineari o non lineari. Il vantaggio dei modelli lineari è l'interpretazione più semplice, ma spesso non sono in grado di fornire previsioni accurate. Al contrario, alcuni metodi non lineari permettono di ottenere previsioni migliori, ma, a causa delle difficoltà di interpretazione, potrebbero essere inadeguati per risolvere un problema di inferenza.

1.2 Metodi parametrici e non parametrici

La maggior parte dei metodi di apprendimento automatico possono essere classificati in parametrici o non parametrici. Tale distinzione è riferita alle ipotesi formulate circa la distribuzione di f. Quando si utilizza un metodo parametrico, in primo luogo, si ipotizza che f abbia una determinata forma funzionale. Successivamente si utilizza un metodo che si

serve dei dati di training per adattare o addestrare il modello. Ad esempio, *f*, ipotizzando che abbia una forma lineare, potrebbe essere espressa con la seguente funzione:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \quad (1.4)$$

In tal caso, è sufficiente stimare il valore dei parametri β_0, β_1,…, β_p con l'utilizzo di diverse tecniche. In generale, i metodi parametrici facilitano la stima di *f* e l'interpretazione del modello, poiché è più semplice stimare i parametri definiti a priori che utilizzare una forma di *f* interamente arbitraria. Inoltre, non richiedono un numero elevato di osservazioni e sono più veloci dal punto di vista computazionale. Uno degli svantaggi è dovuto al fatto che la stima non sarà accurata nei casi in cui il modello scelto si discosti molto dalla vera distribuzione di *f*. L'utilizzo di modelli parametrici flessibili, che possono adattarsi a diverse forme di *f,* aumenterebbe il numero dei coefficienti da stimare e potrebbe determinare un adattamento eccessivo. Al contrario, i metodi non parametrici non ipotizzano esplicitamente la forma funzionale di *f* e, di conseguenza, sono liberi di apprendere qualsiasi forma dai dati di training. Per tale ragione, hanno il vantaggio di potersi adattare a una gamma più ampia di possibili forme per *f* e fornire previsioni più accurate. Poiché i parametri non sono definiti a priori, uno degli svantaggi è legato al numero di osservazioni necessarie per ottenere una stima accurata. Inoltre, il rischio di un adattamento eccessivo è più elevato e in questi casi il modello non produrrebbe stime accurate se applicato a nuove osservazioni non incluse nei dati di training. È evidente che sia presente un trade-off tra accuratezza e interpretabilità dei risultati. Se l'obiettivo è l'inferenza si preferisce utilizzare un modello più interpretabile e restrittivo, poiché un modello più flessibile e complesso renderebbe difficoltoso comprendere la relazione di ogni singolo predittore con la variabile di risposta. Per la stessa ragione si potrebbe pensare che le previsioni più accurate si ottengano utilizzando il metodo più flessibile disponibile, invece, a causa del rischio di adattamento eccesivo, spesso si ottengono migliori prestazioni con un metodo meno flessibile.

1.3 Apprendimento supervisionato e non supervisionato

Nel campo dell'apprendimento automatico la maggior parte delle tecniche possono essere classificate in apprendimento supervisionato e non supervisionato. Il termine *supervisionato* deriva dall'idea che il processo di apprendimento di un algoritmo dal training set può essere immaginato come un insegnante che supervisiona l'intero processo. In questi casi si è a conoscenza della variabile di risposta di ogni esempio dato in input all'algoritmo nella fase di addestramento. Infatti, per ogni osservazione, alle caratteristiche note x_p i = 1, …, n è associato l'output corrispondente y_i. Attraverso le osservazioni del training set viene costruito un modello che ha l'obiettivo di prevedere in maniera accurata l'output per le nuove osservazioni o verificare le relazioni tra gli input e la variabile di risposta. In base alla natura della variabile di risposta è possibile distinguere i problemi di regressione dai problemi di classificazione. Infatti, se l'output è una variabile quantitativa, ad esempio il prezzo di un appartamento o il reddito di una persona, si fa riferimento ai problemi di regressione, invece, nei casi in cui la variabile è qualitativa, perciò può assumere valori in K diverse classi o categorie, si tratta di problemi di classificazione. La scelta del modello di apprendimento automatico è determinata dalla natura della variabile di risposta. Ad esempio, la regressione lineare è utilizzata quando la variabile di risposta è quantitativa, invece, la regressione logistica viene generalmente utilizzata per risolvere problemi di classificazione. Al contrario, nell'apprendimento non supervisionato non è possibile osservare una variabile di risposta che possa supervisionare il processo di apprendimento. L'obiettivo principale è analizzare eventuali relazioni tra le variabili o tra le osservazioni. Alcuni strumenti noti di apprendimento non supervisionato sono l'analisi dei gruppi o clustering e l'analisi delle componenti principali. Tali strumenti spesso trovano gruppi o schemi nascosti all'interno dei dati che un osservatore umano potrebbe non rilevare. Dato un insieme di osservazioni, è possibile utilizzare il clustering per classificare ciascuna osservazione in un gruppo specifico. In teoria, le osservazioni che si che si trovano nello stesso gruppo dovrebbero avere caratteristiche simili. Ad esempio, in uno studio di segmentazione di mercato, è possibile utilizzare il clustering per individuare gruppi omogenei dei potenziali clienti sulla base delle loro caratteristiche note. Nonostante la maggior parte dei problemi di apprendimento possano essere distinti in supervisionato o non supervisionato, in alcune situazioni tale distinzione è meno chiara. Ci si riferisce ad un

problema di apprendimento semi-supervisionato nel caso in cui le misurazioni di risposta sono note solamente per una parte delle osservazioni. Nell'apprendimento semi-supervisionato si utilizza un metodo in grado di includere sia le osservazioni per le quali sono disponibili gli output sia quelle per le quali sono disponibili solamente i predittori.

1.4 Valutazione dell'accuratezza di un modello

Nell'ambito dell'apprendimento automatico non è possibile identificare un modello migliore in assoluto, che avrà prestazioni elevate a prescindere dal data set analizzato. Infatti, un metodo specifico potrebbe funzionare meglio rispetto ad altri su un particolare set di dati, ma potrebbe dare scarsi risultati se applicato ad altri. Al fine di selezionare il modello più adeguato è necessario considerare alcune misure di performance.

1.4.1 Valutazione dei modelli di regressione

La misura più comunemente impiegata nei problemi di regressione è l'errore quadratico medio o *mean squared error* (MSE) dato da:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{1.5}$$

Dove $\hat{f}(x_i)$ è la stima che $\hat{f}$ fornisce per l'*i*-esima osservazione. L'errore quadratico medio sarà piccolo se le stime del modello sono molto vicine rispetto ai valori reali e sarà più grande se le risposte previste per alcune osservazioni differiscono notevolmente da quelle reali. È importante considerare che questo valore MSE viene calcolato utilizzando solo i dati di addestramento sui quali è stato costruito il modello, quindi dovrebbe essere definito come training MSE. In realtà si è maggiormente interessati all'accuratezza del modello quando viene applicato a nuove osservazioni, non incluse nel training set. Ad esempio, se l'obiettivo del modello è prevedere i ricavi di vendita sulla base dei ricavi ottenuti negli anni precedenti non vi è alcun interesse ad ottenere previsioni accurate dei ricavi passati. Per tale ragione l'obiettivo più importante è selezionare il modello che abbia il test MSE più basso tra tutti gli altri modelli possibili. In alcuni casi è possibile calcolare direttamente tale valore poiché si ha a disposizione un set di osservazioni che non sono state utilizzate nella fase di training del modello, ma spesso accade che queste osservazioni non siano disponibili. In queste situazioni si potrebbe utilizzare un metodo che minimizzi il training

MSE, assumendo che sia strettamente correlato al test MSE, ma non vi è alcuna garanzia che il modello con le performance più elevate per il training set ottenga prestazioni simili se applicato ad altre osservazioni. Infatti, alcuni metodi di apprendimento automatico potrebbero comportare un problema di adattamento eccessivo e individuare pattern che sono semplicemente dati dal caso piuttosto che da proprietà reali della funzione ignota *f*. La figura 1.3 mostra un esempio del fenomeno appena descritto.

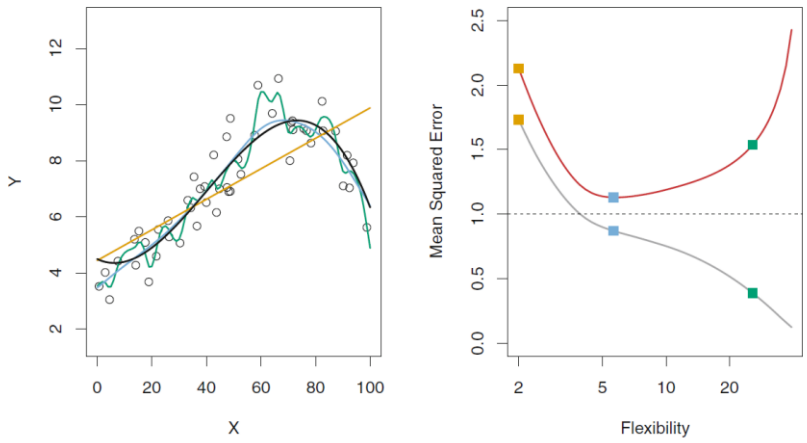


Figura 1.3 A sinistra: in nero dati simulati da f. Sono presentate tre stime di f: una con la regressione lineare (curva arancione) e due con due metodi più flessibili (curve blu e verdi). A destra: è rappresentato l'errore quadratico medio di training (curva grigia), l'errore quadratico medio di test (curva rossa) e l'errore quadratico medio di test minimo rispetto a tutti gli altri metodi (curva tratteggiata). I quadrati rappresentato il training MSE e il test MSE per i tre modelli della parte sinistra. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

Dalla figura 1.3 si può notare che la curva verde è la più flessibile e si adatta molto bene ai dati del training set, rappresentati dai punti bianchi. Tuttavia, non si adatta perfettamente alla vera *f* poiché assume una forma troppo ondulata. Nel grafico a destra le curve grigia e rossa rappresentano il training MSE medio e il test MSE medio in funzione della flessibilità e i quadrati arancione, blu e verde costituiscono i valori del MSE associati alle curve del grafico a sinistra. Il valore minimo che può assumere il test MSE è indicato dalla linea tratteggiata, che rappresenta l'errore non riducibile. Si può osservare che il training MSE diminuisce all'aumentare della flessibilità, perciò sembrerebbe conveniente utilizzare il metodo più flessibile disponibile. Invece, il test MSE ha una diminuzione iniziale, ma ad un certo punto aumenta. Per questa ragione, il metodo più flessibile, rappresentato dalla curva verde, produce un training MSE piccolo e un test MSE piuttosto elevato. Nonostante ci si aspetti quasi sempre che il training MSE sia più piccolo del test MSE, in questa situazione si è in presenza di un problema di adattamento eccessivo (overfitting), poiché un modello meno flessibile produce un test MSE più piccolo. La forma a U osservata nella

curva del test MSE è il risultato di due proprietà concorrenti dei metodi di apprendimento automatico: la varianza e la distorsione. Infatti, il test MSE atteso per un dato valore x_0 è composto da tre quantità fondamentali: la varianza di $\hat{f}(x_0)$, la distorsione (bias) quadratica di $\hat{f}(x_0)$ e la varianza dell'errore ε. Al fine di minimizzare il test MSE atteso è necessario scegliere un metodo che consente di avere contemporaneamente una bassa varianza e una bassa distorsione. La varianza si riferisce al valore con il quale $\hat{f}$ cambierebbe effettuando tale stima usando un diverso data set di training. Poiché i training set vengono utilizzati per adattare il metodo di apprendimento automatico, diversi data set di training comporteranno una diversa $\hat{f}$. Idealmente, la stima di f non dovrebbe variare troppo tra i diversi training set. Tuttavia, se un metodo ha una varianza elevata, piccoli cambiamenti possono comportare grandi cambiamenti in $\hat{f}$. Invece, la distorsione si riferisce all'errore che viene introdotto approssimando un problema reale, che può essere estremamente complesso, con un modello molto più semplice. Ad esempio, la regressione lineare presuppone l'esistenza di una relazione lineare tra Y e X. È improbabile che qualsiasi problema della vita reale abbia davvero una relazione lineare così semplice, e quindi eseguire la regressione lineare si tradurrà senza dubbio in qualche distorsione nella stima di f. In generale, i metodi statistici più flessibili comportano una minore distorsione, ma allo stesso tempo una varianza più elevata. Per tale ragione, nella scelta del modello di apprendimento automatico è necessario considerare l'esistenza del trade off tra distorsione e varianza. Il tasso relativo di variazione di queste due quantità determinerà se il test MSE aumenterà o diminuirà in funzione della flessibilità. Nell'esempio della figura 1.3, inizialmente la distorsione diminuisce più rapidamente dell'aumento della varianza, perciò il test MSE diminuisce. Tuttavia, ad un certo punto l'aumento della flessibilità ha un impatto superiore sulla varianza rispetto alla distorsione e il test MSE aumenta.

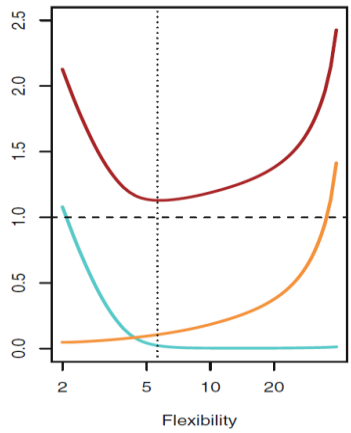


Figura 1.4. Distorsione al quadrato (curva blu), la varianza (curva arancione), l'errore non riducibile (linea tratteggiata) e il test MSE (curva rossa) per il dataset presentato nella figura 1.2. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

1.4.2 Valutazione dei modelli di classificazione

Nei problemi di classificazione l'approccio più comune per valutare l'accuratezza della stima di *f* è il tasso di errore, cioè la proporzione di osservazioni classificate erroneamente dal modello. Anche nel contesto della classificazione si è maggiormente interessati alle prestazioni del modello che risultano dalla sua applicazione a nuove osservazioni, non utilizzate nella fase di addestramento. Per tale ragione una delle misure utilizzate è il tasso di errore di test. Il tasso di errore è minimizzato dal *classificatore di Bayes* che assegna ogni osservazione alla classe più verosimile, dati i valori dei predittori. Infatti, una determinata osservazione di test è assegnata a una determinata classe *j* a seconda della probabilità condizionata Y = *j* dato il valore dei predittori osservati X.

Dal tasso di errore si può definire l'accuratezza di un modello, che indica la frazione di osservazioni classificate correttamente. Utilizzare solamente il tasso di errore e l'accuratezza per la valutazione del modello potrebbe comportare a delle conclusioni sbagliate in determinati casi. Questo problema è evidente nelle situazioni in cui vi è un forte sbilanciamento tra le classi di risposta del dataset. Ad esempio, si supponga di voler rilevare le frodi in un dataset di transazioni finanziarie, costituito al 99% da transazioni sicure. Se si utilizzasse un modello che classifica ogni nuova transazione come sicura si otterrebbe un tasso di errore pari al 1%. Ovviamente non si tratterebbe di un modello accurato poiché non riuscirebbe a identificare le frodi. Pertanto, è necessario utilizzare ulteriori metriche, che possono essere ottenute a partire dai valori della *matrice di confusione*. La matrice di

confusione mostra le previsioni corrette ed errate su ciascuna classe. Se la classe di risposta può assumere solamente due valori è possibile rappresentarla tramite la seguente tabella:

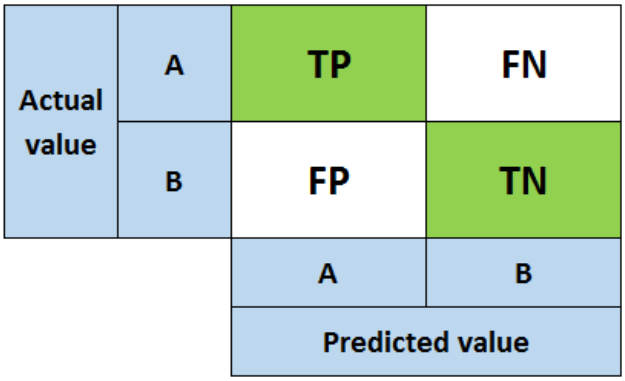


Figura 1.5 Matrice di confusione 2x2 di una classificazione binaria

- TP è il numero dei casi positivi classificati correttamente come positivi;
- FP è il numero dei casi negativi classificati erroneamente come positivi;
- TN è il numero dei casi negativi classificati correttamente come negativi;
- FN è il numero dei casi positivi classificati erroneamente come negativi.

Dai valori della matrice di confusione è possibile ricavare le seguenti metriche di prestazione:

- Specificità: indica la percentuale dei casi negativi classificati correttamente come negativi. Una specificità elevata indica che il classificatore rileva correttamente le osservazioni negative.

$$Specificità = \frac{TN}{TN + FP}$$

- Sensibilità: indica la capacità di un classificatore di rilevare correttamente tutte le osservazioni positive. Per ogni classe è definito come il rapporto tra i veri positivi e la somma dei veri positivi e dei falsi negativi.

$$Sensibilità = \frac{TP}{TP + FN}$$

- Precisione: indica la frazione dei veri positivi rispetto a tutti i risultati positivi.

$$Precisione = \frac{TP}{TP + FP}$$

- F1: combina sensibilità e precisione, che spesso variano in maniera inversamente proporzionale, e può essere definita come la loro media armonica. β è un

parametro che esprime l'importanza relativa della precisione e della sensibilità, che generalmente è pari a 1.

$$F1 = \frac{(1 + \beta^2) * Sensibilità * Precisione}{\beta^2 * Sensibilità * Precisione}$$

Al fine di assegnare una determinata osservazione alla classe di appartenenza viene utilizzato un valore soglia. Ad esempio, è possibile stabilire che nei casi in cui la probabilità (calcolata dal modello di classificazione) di appartenere alla classe positiva sia uguale o superiore al 50%, queste osservazioni siano etichettate come positive. Il valore soglia del 50% svolge un ruolo fondamentale, poiché prendendo in considerazione valori diversi si otterrebbero diverse matrici di confusione e, di conseguenza, diverse metriche di prestazione.

È possibile utilizzare uno strumento che riassume le prestazioni del modello a diversi valori di soglia combinando tutte le matrici di confusione possibili: la curva delle caratteristiche operative del ricevitore (ROC).[3] La curva ROC è una tecnica standard per riepilogare le prestazioni del classificatore su una serie di trade-off tra i tassi di veri positivi e falsi positivi.[4] L'asse X della curva ROC è il tasso di veri positivi (sensibilità) e l'asse y della curva ROC è il tasso di falsi positivi (1 – specificità).

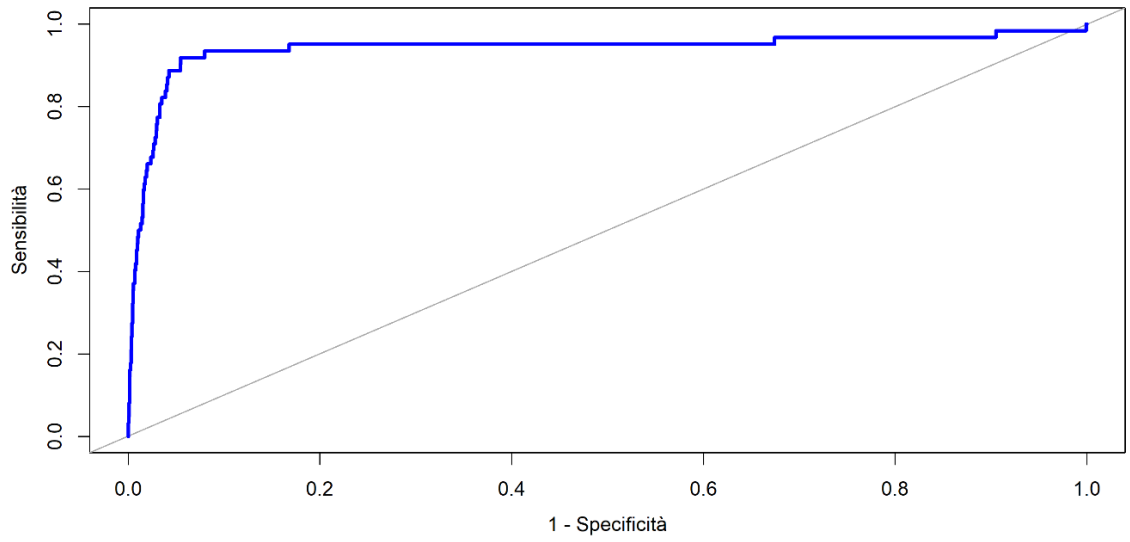


Figura 1.6 Esempio di curva ROC

[3] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
[4] Swets, J. (1988). Measuring the Accuracy of Diagnostic Systems.

Ci sono diversi punti importanti in una curva ROC:

- (0,0) che rappresenta la situazione in cui non vi è alcuna classificazione positiva; in questo modo non ci sono errori relativi ai falsi positivi, ma non vengono rilevati neanche i veri positivi.
- (1,1) che rappresenta la situazione opposta in cui vi è sempre una classificazione positiva; in questo caso non ci sono falsi negativi, ma chiaramente non vengono rilevati neanche i veri negativi.
- (0,1) che rappresenta la classificazione perfetta, in cui il modello ha la capacità di classificare correttamente tutte le osservazioni.

La metrica utilizzata per la valutazione delle prestazioni di una curva ROC è l'area sotto la curva (AUC).[5] L'AUC fondamentalmente aggrega le prestazioni del modello a tutti i valori di soglia. Il valore dell'AUC è compreso tra 0 e 1. Il miglior valore possibile di AUC è 1 e indica un classificatore perfetto, in cui tutti i casi positivi sono classificati correttamente e i casi negativi non sono classificati erroneamente come positivi.[6] Inoltre, la diagonale rappresentata nel grafico rappresenta il caso del classificatore casuale, in cui l'AUC è pari a 0.5. Le curve ROC possono essere utilizzate per confrontare due o più modelli di classificazione, ma anche al fine di selezionare il valore soglia per la probabilità che garantisce i valori di specificità e sensibilità desiderati e richiesti dal problema.

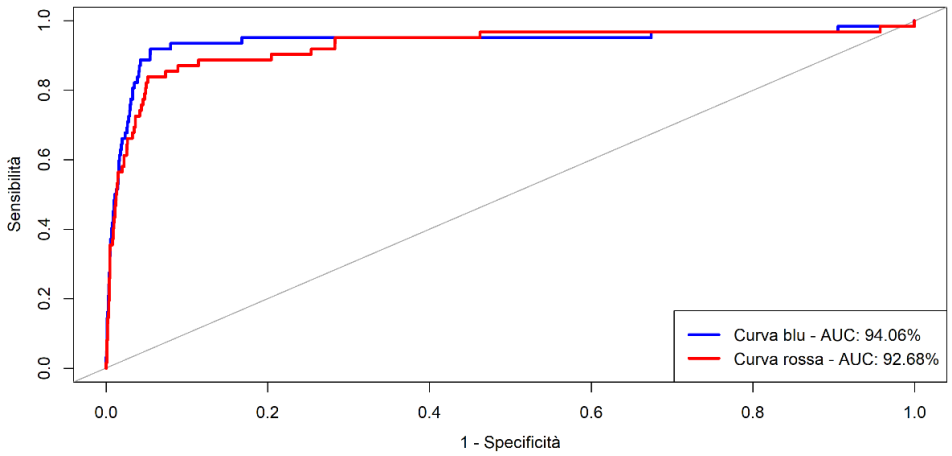


Figura 1.7 Esempio di confronto tra due Curve ROC e relativo valore di AUC

[5] Bradley, A. P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition.
[6] Chawla, N. W. (2005). Data mining for imbalanced datasets: an overview.

2. I modelli di classificazione

La classificazione è una tecnica utilizzata nell'apprendimento supervisionato, il cui obiettivo è riuscire a prevedere le etichette delle classi per i nuovi dati, sulla base delle precedenti osservazioni. Pertanto, classificare un'osservazione significa prevedere la corrispondente variabile di risposta qualitativa, in quanto l'osservazione viene assegnata ad una categoria o classe. Esistono molte tecniche di classificazione utilizzabili per prevedere una risposta qualitativa. Di seguito saranno descritti tre metodi: la regressione logistica, l'analisi discriminante lineare e le support vector machines.

2.1 Regressione logistica

La regressione logistica è utilizzata per prevedere la probabilità di occorrenza di un evento attraverso l'utilizzo di una funzione logistica. Si applica per la risoluzione di problemi di classificazione, in particolare quando la variabile dipendente Y è dicotomica. Si differenzia dalla regressione lineare, poiché per quest'ultima si ipotizza una distribuzione normale di Y, mentre se Y è dicotomica la sua distribuzione è binomiale. La funzione logistica utilizzata per la stima di Y può essere definita da:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+ \beta_2 X_2+ \dots + \beta_p X_p}}{1+ e^{\beta_0+\beta_1 X_1+ \beta_2 X_2+ \dots + \beta_p X_p}} \qquad (2.1)$$

La stima di Y nella regressione lineare varia da -∞ a +∞, mentre nella regressione logistica varia tra 0 e 1, trattandosi di una probabilità. Infatti, la funzione logistica produce sempre una curva a S delimitata nell'asse Y da 0 e 1, come illustrato nella figura 2.1.

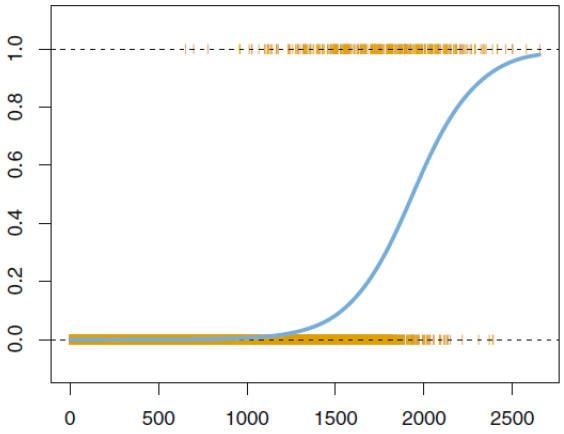


Figura 2.3. Probabilità previste per la variabile Y usando la regressione logistica. Tutti i valori di probabilità sono compresi tra 0 e 1. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

La figura 2.1 rappresenta un modello piuttosto semplice, in cui è utilizzata solamente una variabile indipendente X per prevedere la variabile dipendente Y. In questo caso si osserva che per valori molto alti di X (o molto bassi se la relazione è negativa) il valore in Y è molto vicino ad 1 e non deve superare tale limite. La stessa situazione è visibile in prossimità dello 0. Talvolta può essere utile sostituire la probabilità della (2.1) con l'odds corrispondente. L'odds esprime il rapporto tra la probabilità p di un evento e la probabilità 1 - p che tale evento non accada. È possibile riscrivere la (2.1) in questo modo:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X_1+ \beta_2 X_2+ \dots + \beta_p X_p} \tag{2.2}$$

La quantità $\frac{p(X)}{1-p(X)}$ è chiamata odds e può assumere qualunque valore tra 0 e +∞. Valori di odds prossimi allo 0 indicano probabilità molto basse, prossimi a +∞ indicano probabilità molto alte.[7] Inoltre, considerando il logaritmo di entrambi i membri della (2.2) si ottiene:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \tag{2.3}$$

La parte destra di tale equazione viene chiamata log-odds o logit. L'aumento di X_1, di un'unità modifica la probabilità del log-odds di β_1. Al contrario della regressione lineare, in cui una variazione di β_1 corrisponderebbe alla variazione di $p(X)$ associata all'aumento di un'unita di X_1, nella regressione logistica tale aumento dipende dal valore corrente di X_1. Indipendentemente dal valore corrente è possibile affermare che l'aumento di X_1 sarà associato all'aumento di $p(X)$ se β_1 è positivo, invece, si verificherà una diminuzione di $p(X)$ se β_1 è negativo. I coefficienti $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ non sono noti, perciò vengono stimati utilizzando il metodo della massima verosimiglianza. Tale metodo stima i coefficienti del modello in modo da massimizzare la funzione che indica quanto è probabile ottenere il valore atteso di Y dati i valori delle variabili indipendenti. Si supponga che Y = 1 identifichi lo stato di insolvenza e Y = 0 identifichi gli individui solventi. I valori di $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \dots, \widehat{\beta_p}$ devono essere tali che introdotti nel modello della (2.1) si ottenga un valore vicino a uno per tutti gli individui insolventi e un numero vicino a zero per tutti gli individui solventi.

[7] James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

2.2 Analisi discriminante lineare

L'analisi discriminante lineare è un approccio alternativo alla regressione logistica che generalmente è preferibile utilizzare in alcune situazioni:

- Le classi di Y sono ben separate dai predittori X, in questo caso le stime per il modello di regressione logistica sono instabili;
- *n* è piccolo e la distribuzione dei predittori X è approssimativamente normale in ciascuna delle classi;
- La variabile di risposta Y ha più di due classi.

A differenza della regressione logistica, che modella la probabilità condizionata di Y dati i predittori X attraverso la funzione logistica, l'approccio dell'analisi discriminante è indiretto e si serve del teorema di Bayes per stimare $P(Y = k|X = x)$.
Si supponga di voler classificare un'osservazione in una delle K classi possibili, con K ≥ 2. La probabilità a priori che una determinata osservazione appartenga alla classe k-esima di Y può essere definita π_k . Sia $f_k(x) \equiv P(X = x \mid Y = k)$ la funzione di densità di X per un'osservazione della classe k-esima. Il teorema di Bayes afferma che[8]:

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{k} \pi_l f_l(x)} \quad (2.4)$$

$P(Y = k|X = x)$ indica la probabilità a posteriori che un'osservazione appartenga alla classe k dato il valore dei predittori $X = x$. La probabilità a priori π_k può essere facilmente stimata utilizzando le frequenze relative di ciascuna classe k, perciò il problema si riduce alla stima di $f_k(x)$. L'analisi discriminante lineare, attraverso alcune assunzioni di base, stima $f_k(x)$ al fine di sviluppare un classificatore che approssimi il classificatore di Bayes. L'assunzione di base in presenza di un solo predittore X è quella per cui $f_k(x)$ assuma una forma *normale*. La funzione di densità di una normale univariata è:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right) \quad (2.5)$$

[8] James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

Dove μ_k e σ_k^2 sono rispettivamente la media e la varianza per la classe k-esima. Inoltre, si assume che la varianza sia uguale per tutte le classi, perciò può essere indicata con σ^2.
La probabilità a posteriori definita dal teorema di Bayes può essere riscritta come:

$$P(Y = k|X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \quad (2.6)$$

Come indicato nel capitolo 1, il classificatore di Bayes assegna ogni osservazione alla classe più verosimile, per la quale la probabilità a posteriori è massima. Rielaborando l'espressione della (2.6) si ottiene che tale probabilità è massima quando:

$$\delta_k(x) = x * \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \quad (2.7)$$

è massima. Poiché non è possibile conoscere direttamente alcuni dei valori inseriti nella (2.7), l'analisi discriminante lineare approssima il classificatore di Bayes attraverso le stime di π_k, μ_k e σ^2.
La stima di μ_k è data dalla media di tutte le osservazioni del training set della classe k-esima, mentre la stima di σ^2 è la media ponderata delle varianze campionarie per ciascuna delle K classi. In generale è possibile osservare direttamente π_k per ciascuna classe, ma in assenza di informazioni si utilizza il valore stimato, dato dal rapporto tra il numero di osservazioni del training set della classe k-esima e il numero di osservazioni totali del training set. I parametri stimati sono sostituiti alle formule precedenti per ottenere le stime della probabilità di appartenenza ad una determinata classe e per la classificazione. La definizione lineare derivata dal fatto che la stima di $\delta_k(x)$ è una funzione lineare di x.
Nel caso in cui vi siano più predittori X l'assunzione di base dell'analisi discriminante lineare è differente. In queste situazioni si assume che le osservazioni di ciascuna classe siano estratte da una distribuzione normale multivariata $X \sim N(\mu_k, \sum)$, in cui μ_k è il vettore delle medie specifico peer ciascuna classe e Σ è la matrice di covarianza comune a tutte le classi. Dopo aver effettuato la stima dei parametri μ_k, Σ e π_k le probabilità condizionate sono stimate ricorrendo al teorema di Bayes e utilizzate per la classificazione delle osservazioni. Nella figura 2.2 è possibile osservare un esempio con tre classi estratte da una distribuzione normale multivariata con due predittori X.

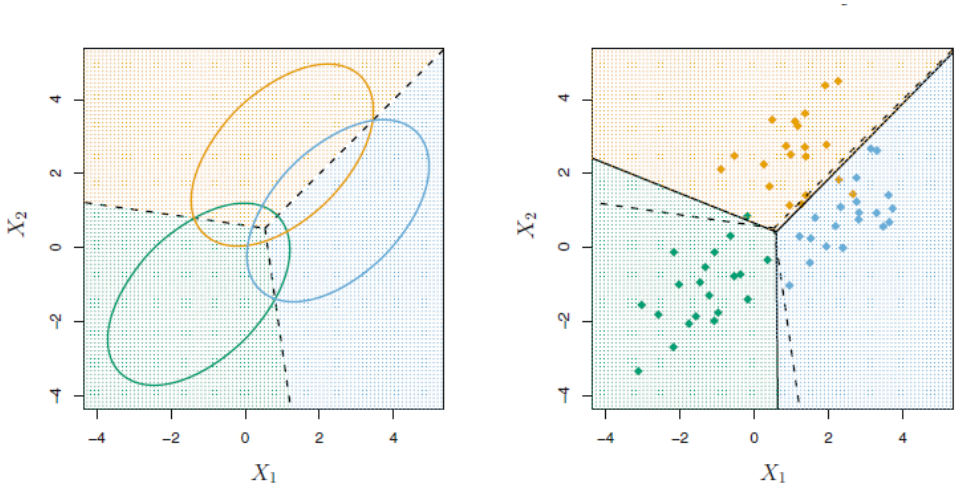


Figura 4.2. Un esempio con tre classi. Le osservazioni di ciascuna classe sono estratte da una distribuzione normale multivariata con due predittori X, con un vettore delle medie specifico della classe e una matrice di covarianza comune. A sinistra: vengono visualizzate le ellissi che contengono il 95% della probabilità per ciascuna delle tre classi. Le linee tratteggiate indicano i confini decisionali di Bayes. A destra: sono state generate 20 osservazioni da ciascuna classe. I confini di decisione dell'analisi discriminante lineare sono indicati dalle linee continue, mentre i confini di decisione di Bayes dalle linee tratteggiate. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

Si può notare che i confini di decisione di Bayes (linee tratteggiate) e i corrispondenti confini di decisione dell'analisi discriminante lineare (linee continue nel grafico a destra) dividono lo spazio in tre aree. Un'osservazione viene classificata in base all'area in cui si trova.

Per confrontare la regressione logistica e l'analisi discriminante lineare si può considerare per semplicità il caso con un solo predittore e due classi. Si indichi con $p_1(x)$ la probabilità condizionata che Y appartenga alla classe 1 dato X, nell'analisi discriminante lineare c_0 e c_1 sono funzioni di μ_1, μ_2 e σ^2 e il logaritmo dell'odds è dato da ha:

$$\log \frac{p_1(x)}{1-p_1(x)} = c_0 + c_1 x \tag{2.8}$$

Invece, nella regressione logistica si ha:

$$\log \frac{p_1(x)}{1-p_1(x)} = \beta_0 + \beta_1 x \tag{2.9}$$

Entrambe le formule sono funzioni lineare di x, mentre si differenziano poiché i coefficienti della regressione logistica sono stimati usando la massima verosimiglianza, mentre nell'analisi discriminante lineare sono calcolati usando la media e la varianza stimate da una distribuzione normale. Data la somiglianza tra i due metodi, le prestazioni sono spesso simili, tuttavia l'analisi discriminante lineare raggiunge performance migliori se l'assunzione di normalità è soddisfatta, mentre accade il contrario se la distribuzione reale è lontana da quella normale.

2.3 Support vector machines

Le support vector machines sono un insieme di tecniche di apprendimento supervisonato generalmente utilizzate per la risoluzione di problemi di classificazione binaria. Al fine di descrivere il funzionamento di tali tecniche è necessario introdurre il concetto di iperpiano a margine massimo. In uno spazio p – dimensionale, un iperpiano è un sottospazio planare affine di dimensione $p - 1$[9]. Nel caso più semplice, di due dimensioni, un iperpiano è una linea, invece, in tre dimensioni può essere rappresentato da un piano. In un esempio di classificazione binaria si può supporre che sia possibile costruire un iperpiano che separi perfettamente le osservazioni del training set in base alla classe di appartenenza. Nella figura 2.3 il dataset è linearmente separabile e si può notare che esiste un numero infinito di iperpiani che divide le due classi di osservazioni. Per tale ragione, ciascun iperpiano della figura 2.3 è definito *iperpiano separante* e permette di classificare un'osservazione di test in base a quale lato dell'iperpiano si trova.

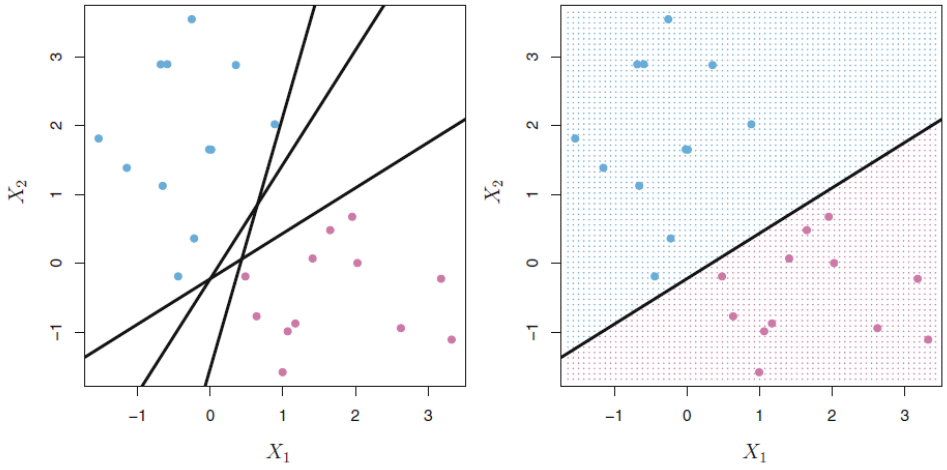


Figura 2.5. A sinistra: due classi di osservazioni, in blu e in viola, ognuna delle quali ha misure su due variabili. Le linee nere rappresentano tre dei possibili iperpiani separanti. A destra: l'iperpiano separante (linea nera) suddivide le osservazioni in due classi. Le osservazioni nella griglia blu sono state assegnate alla classe blu, invece, le osservazioni nella griglia viola sono state assegnate alla classe viola. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

Se le osservazioni della classe blu sono etichettate come $y_i = 1$ e quelle della classe viola con $y_i = -1$ un iperpiano separante ha la proprietà per cui:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) > 0 \qquad (2.10)$$

[9] James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

per tutti gli $i = 1, \dots, n$. L'osservazione di test perciò viene classificata sulla base del segno della (2.10). Se il risultato è positivo l'osservazione viene assegnata alla classe 1, invece, se è negativo viene assegnata alla classe -1. Inoltre, dal valore dell'osservazione è possibile definire il grado di certezza di tale assegnazione. Infatti, se la (2.10) ha un valore lontano da 0, positivo o negativo, vi è una ragionevole certezza che l'osservazione appartenga alla classe assegnata. Il problema è selezionare l'iperpiano separante ottimale, ossia quello che permette di ottenere le migliori performance nella classificazione delle nuove osservazioni. Una scelta naturale è l'*iperpiano a margine massimo*, che è l'iperpiano separante più lontano dalle osservazioni utilizzate per l'addestramento del modello. Il margine è la distanza perpendicolare minima di punti delle due classi nel training set da un determinato iperpiano. L'iperpiano a margine massimo è l'iperpiano separante per il quale il margine è maggiore ed è rappresentato nella figura 2.4. Si può notare che tale iperpiano comporta una maggiore distanza minima tra le osservazioni, rispetto all'iperpiano di separazione rappresentato nell'immagine a destra della figura 2.3.

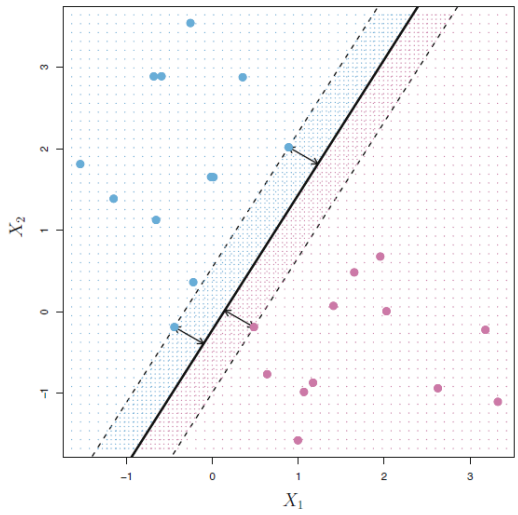


Figura 2.6 L'iperpiano a margine massimo è rappresentato dalla linea continua nera. Il margine è la distanza dalla linea continua a una delle linee tratteggiate. I due punti blu e il punto viola posizionati sulle linee tratteggiate sono i vettori di supporto. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

Dalla figura 2.4 si notano tre osservazioni equidistanti dall'iperpiano a margine massimo e posizionate lungo le linee tratteggiate che indicano la larghezza del margine. Queste osservazioni sono i *vettori di supporto*, così definiti poiché se venissero spostati anche leggermente determinerebbero uno spostamento dell'iperpiano a margine massimo. Al contrario, le altre osservazioni non esercitano una tale influenza, a meno che non si verifichi uno spostamento oltre il confine determinato dal margine.

In alcune situazioni le osservazioni sono disposte in maniera tale da non poter essere separabili da un iperpiano, perciò il dataset viene definito come *non separabile*. Inoltre, il classificatore basato sull'iperpiano a margine massimo potrebbe causare un adattamento eccessivo al training set, poiché, come mostrato nella figura 2.5, è estremamente sensibile a un cambiamento in una singola osservazione.

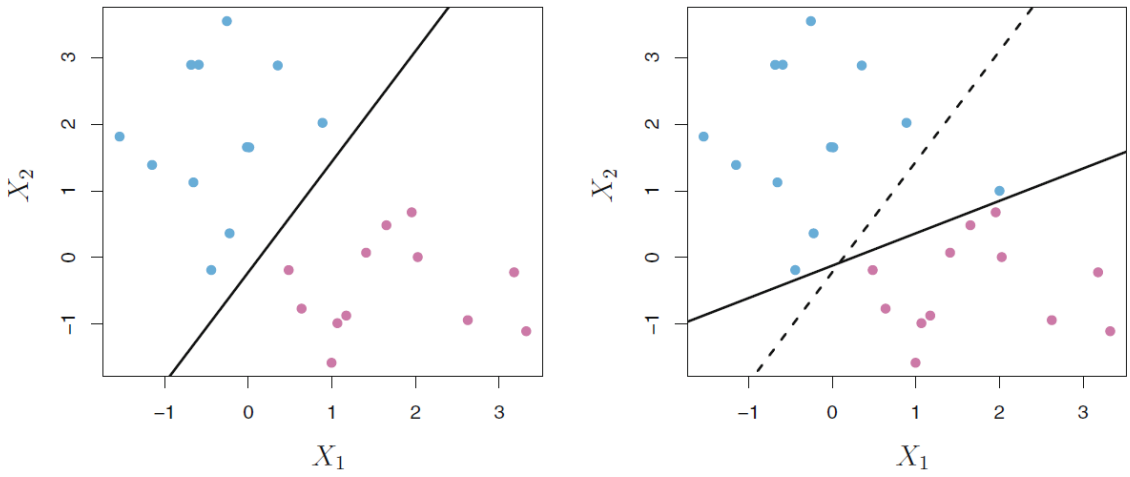


Figura 2.7. A sinistra: due classi di osservazioni in blu e in viole sono separate dall'iperpiano a margine massimo (linea nera continua). A destra: alle osservazioni dell'immagine a sinistra è stata aggiunta un'osservazione che porta a uno spostamento significativo dell'iperpiano. La linea tratteggiata indica l'iperpiano a margine massimo utilizzato in precedenza, la linea continua rappresenta l'iperpiano a margine massimo dopo aver aggiunto una sola osservazione. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

Per queste ragioni, potrebbe essere preferibile utilizzare un altro classificatore basato su un iperpiano che separa erroneamente alcune osservazioni del training set, ma raggiunge prestazioni migliori nella classificazione delle nuove osservazioni: il classificatore *soft margin.*

Il classificatore soft margin consente che alcune osservazioni si trovino sul lato errato del margine o dell'iperpiano stesso, tramite l'utilizzo delle variabili di compensazione ϵ_i per tutti gli $i = 1, \dots, n$ e del parametro di tuning C. Tali variabili ϵ_i indicano l'entità dello spostamento di una singola osservazione relativamente all'iperpiano e al margine: se $\epsilon_i = 0$, l'osservazione si trova sul lato corretto del margine; se $\epsilon_i > 0$ si trova dalla parte sbagliata del margine, invece, se $\epsilon_i > 1$ si trova dalla parte sbagliata dell'iperpiano. Il parametro di tuning C esprime il valore massimo della somma delle ϵ_i, perciò determina il numero e la gravità delle violazioni del margine tollerate e viene trattato come un parametro di ottimizzazione. In questa tecnica di apprendimento automatico C controlla il trade-off tra distorsione e varianza introdotto nel capitolo 1. Quando il parametro C è

piccolo il margine viene violato raramente, perciò il classificatore si adatta alle osservazioni di training e determina una distorsione bassa ma una varianza elevata; quando il parametro C è più elevato sono tollerate più violazioni del margine, perciò il classificatore è maggiormente flessibile. La flessibilità permette di ottenere una varianza inferiore, ma potrebbe causare una distorsione più elevata.

Nella figura 2.6 è possibile osservare due esempi del classificatore soft margin. In entrambi i casi la maggior parte delle osservazioni si trovano sul lato corretto del margine, ma si notano alcune eccezioni. Nell'immagine a sinistra le osservazioni 1 e 8 si trovano sul lato sbagliato del margine, invece, nell'immagine a destra sono state aggiunte le osservazioni 11 e 12 che si trovano sul lato sbagliato dell'iperpiano e del margine.

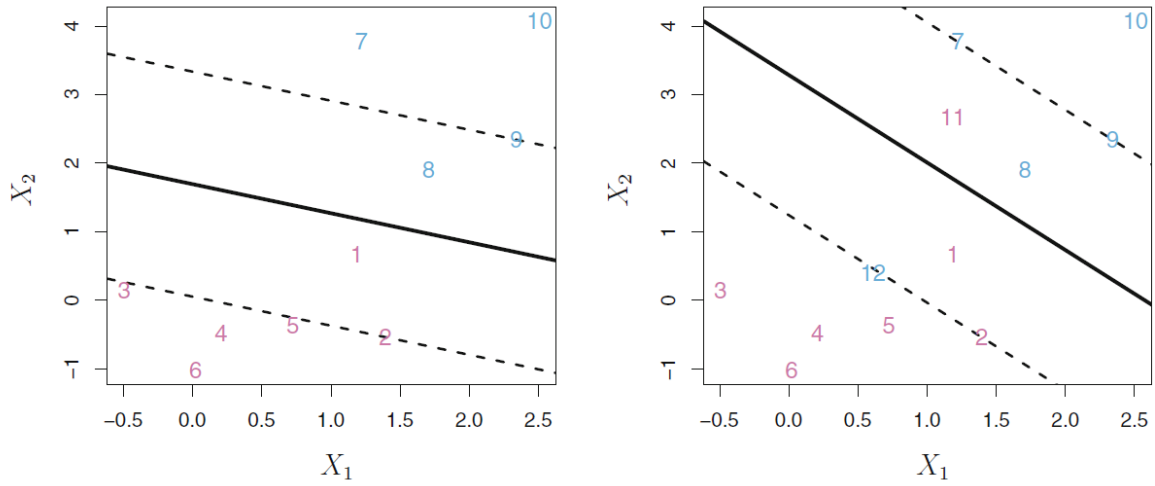


Figura 2.6. A sinistra: un esempio di classificatore soft margin in cui alcune osservazioni si trovano sul lato sbagliato del margine. Nessuna osservazione si trova sul lato sbagliato dell'iperpiano. A destra: sono stati inserite due osservazioni aggiuntive rispetto all'immagine a sinistra. In questo caso sono presenti due osservazioni sul lato sbagliato dell'iperpiano e del margine.
Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

I vettori di supporto per il classificatore soft margin, oltre le osservazioni posizionate direttamente sul margine, sono i punti che si trovano sul lato sbagliato del margine per la loro classe. Tale considerazione spiega il motivo per il quale il parametro C controlla il trade-off tra distorsione e varianza. All'aumentare del parametro C, aumenta anche il numero dei vettori di supporto che determinano la posizione dell'iperpiano, perciò il classificatore ha una varianza bassa, ma una distorsione potenzialmente elevata.

La proprietà del classificatore soft margin di costruire una regola decisionale sulla base di un sottoinsieme delle osservazioni del training set, osservata anche nel classificatore a margine massimo, distingue tali tecniche dall'analisi discriminante lineare. Infatti, la regola

di classificazione dell'analisi discriminante lineare dipende dalla media di tutte le osservazioni di ciascuna classe e dalla matrice di covarianza all'interno della classe calcolata utilizzando tutte le osservazioni. [10]

Nella pratica il classificatore soft margin è raramente utilizzabile, poiché il confine tra le due classi non è lineare. Un esempio è rappresentato nella figura 2.7, in cui le osservazioni non sono separabili linearmente e l'applicazione del classificatore soft margin determina scarse prestazioni.

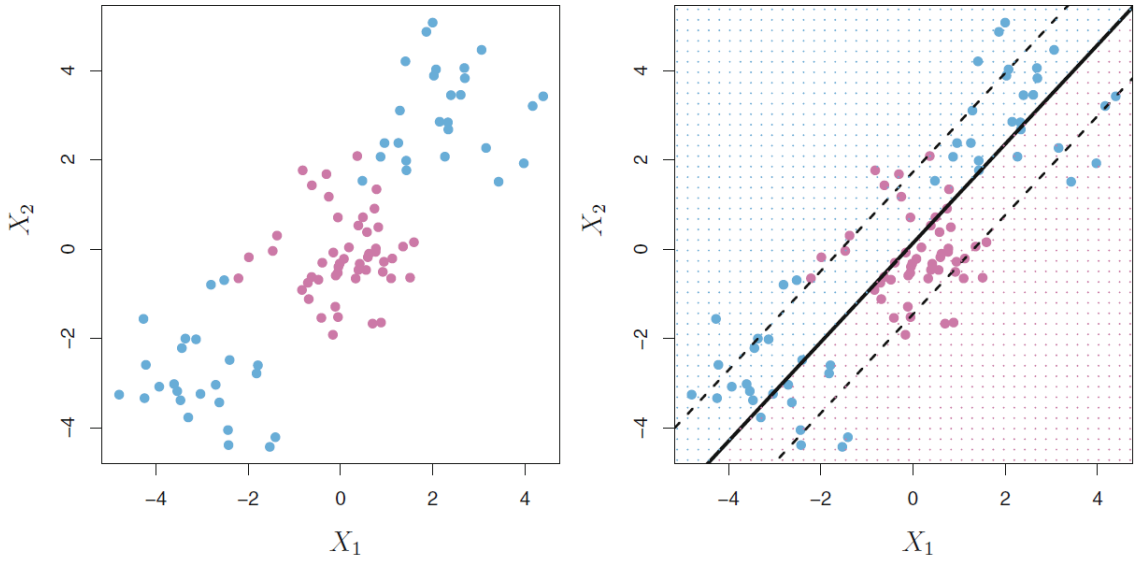


Figura 2.7. A sinistra le osservazioni si dividono nelle classi blu e viola, con un confine non lineare tra loro. A destra: il classificatore soft margin non riesce a stabilire correttamente un confine lineare tra le due classi di osservazioni. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

In tali situazioni è possibile migliorare le prestazioni attraverso la support vector machine, un'estensione del classificatore soft margin, che utilizza i *kernel* per ampliare lo spazio delle variabili in un modo specifico.

Al fine di introdurre il funzionamento dei kernel, innanzitutto è necessario definire il classificatore *soft margin* attraverso la funzione:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \qquad (2.11)$$

In cui è presente un parametro α_i per ogni osservazione del training set. Inoltre, α_i è diverso da zero solamente per i vettori di supporto, perciò è necessario calcolare solamente

[10] James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

il prodotto scalare per questo sottoinsieme di osservazioni. Il support vector machine sostituisce il prodotto scalare della (2.11) con una sua generalizzazione, chiamata appunto kernel. Un kernel è una funzione che quantifica la somiglianza di due osservazioni. Nel classificatore soft margin il kernel è lineare. Ulteriori kernel utilizzati per la classificazione di dataset non lineari sono il kernel polinomiale e il kernel radiale. Il kernel polinomiale permette di ottenere un confine decisionale più flessibile e consiste in un adattamento del classificatore soft margin in uno spazio di dimensioni superiori, rappresentato nell'immagine a sinistra della figura 2.8.

Invece, il kernel radiale, applicato nell'immagine a destra della figura 2.8, è definito da:

$$K\ (x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{ij'})^2\right) \qquad (2.12)$$

In cui γ è una costante positiva. All'aumentare d tale costante il modello diventa più flessibile, perciò aumenta il rischio di un adattamento eccessivo alle osservazioni del training set. Nell'applicazione del kernel radiale, è interessante notare che all'aumentare della distanza Euclidea tra una nuova osservazione e un'osservazione del training set, quest'ultima non avrà alcun effetto sulla classificazione.

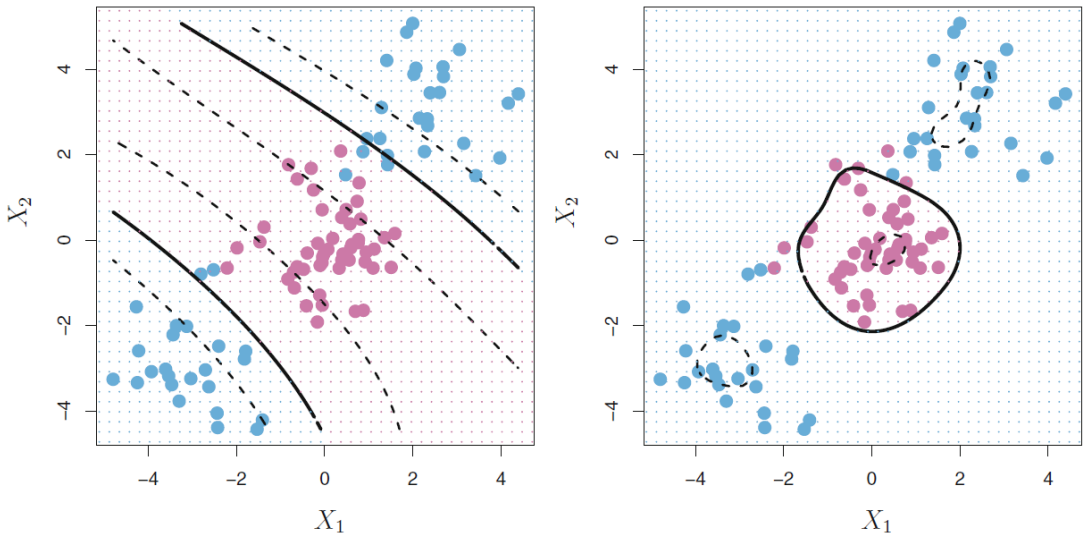


Figura 2.8 A sinistra: un kernel polinomiale con d = 3 viene applicato alle osservazioni della figura 2.7. Si può notare una classificazione più appropriata. A destra: un kernel radiale applicato alle medesime osservazioni. Anche in questo caso le prestazioni sono migliori rispetto al classificatore soft margin. Fonte: James, G. et al. (2013). An Introduction to Statistical Learning: With Applications in R.

3. Il problema dei dataset sbilanciati

Spesso nei problemi di classificazione, la distribuzione delle classi di risposta è piuttosto sbilanciata. Un set di dati è sbilanciato se le categorie di classificazione sono rappresentate da un numero di osservazioni estremamente diverso. Nella classificazione binaria l'etichetta che ha il numero di osservazioni più elevato viene definita classe di maggioranza, mentre l'altra etichetta è definita classe di minoranza.

Lo squilibrio tra le classi ha un impatto negativo sia nella stima che nella valutazione dell'accuratezza del modello. Infatti, il modello tende a focalizzarsi sulla classe di risposta prevalente e ignorare gli eventi rari, che solitamente rappresentano il concetto di interesse. Ad esempio, se l'obiettivo di una banca è rilevare le transazioni non sicure, sarà necessario tenere in considerazione che le frodi sono eventi rari e la classe prevalente è costituita dalle transazioni sicure.

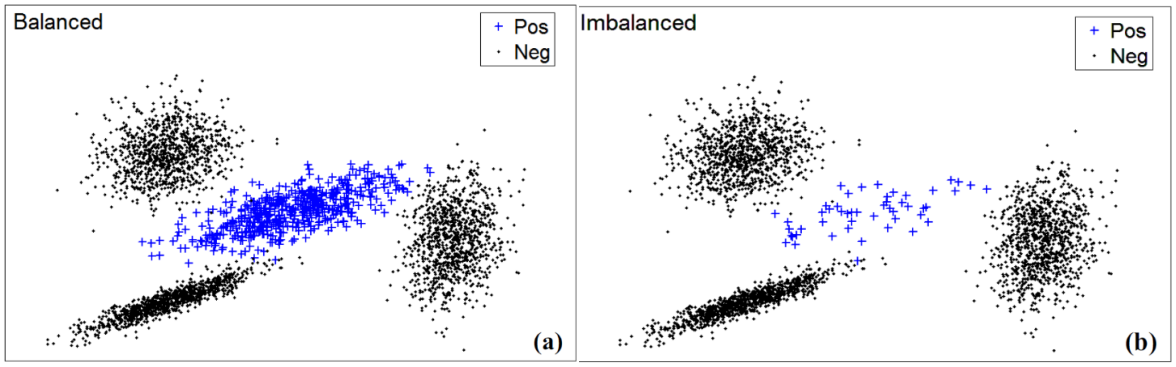


Figura 3.1 A sinistra: un esempio di dataset bilanciato. La classe positiva (blu) ha una numerosità simile alla classe negativa (nera). A destra: un esempio di dataset sbilanciato. La classe positiva (blu) è rappresentata da un numero di osservazioni estremamente piccolo rispetto alla classe negativa (nera). Fonte: Cao, H., Li, X. L., Woon, Y. K., & Ng, S. K., (2011) SPO: Structure preserving oversampling for imbalanced time series classification.

Lo sbilanciamento di un dataset può essere legato a diverse ragioni. La maggior parte di tali squilibri sono indicati come intrinseci poiché sono dovuti alla natura del dataset, come nel caso delle frodi. Tuttavia, in alcuni casi lo sbilanciamento è dovuto a fattori esterni, ad esempio all'eccessiva onerosità nella fase di raccolta di informazioni legate a una particolare classe. Quando lo sbilanciamento non è dovuto alla natura del dataset, si parla di squilibri estrinseci. [11]

Oltre a squilibrio intrinseco ed estrinseco, è rilevante la differenza tra squilibrio relativo e squilibrio dovuto a rarità assoluta. La differenza tra i due casi è legata alle dimensioni del dataset. Infatti, nei dataset di dimensioni elevate, nonostante lo sbilanciamento, la classe

[11] He, H., Garcia, E. A., (2009). Learning from Imbalanced Data.

minoritaria potrebbe essere caratterizzata da una numerosità molto elevata. In tal caso si parla di squilibrio relativo, poiché la frequenza è tale da consentire l'utilizzo dei metodi più comuni di apprendimento automatico e raggiungere prestazioni adeguate. Al contrario, si parla di squilibrio assoluto quando la frequenza della classe di minoranza è bassa, nonostante l'elevata numerosità del dataset.

Le prestazioni dei metodi di classificazione descritti nel capitolo precedente risentono dello sbilanciamento delle classi di risposta per diversi motivi. Innanzitutto, assumono che ci sia un'equa distribuzione dei dati per tutte le classi e che gli errori provenienti dalle diverse classi abbiano lo stesso peso. Inoltre, il loro obiettivo è di minimizzare l'errore globale al quale la classe minoritaria contribuisce poco.

La regressione logistica, ad esempio, non è consigliabile quando le classi sono sbilanciate, poiché la probabilità condizionata della classe rara viene sottostimata.[12] Invece, l'analisi discriminante lineare assume una matrice di covarianza comune tra le classi di risposta, che viene stimata dalle matrici campionarie in ciascuna delle classi. In caso di sbilanciamento la matrice di covarianza sarebbe prevalentemente influenzata dalla classe più numerosa e le previsioni del modello potrebbero essere distorte. [13]

Analogamente agli altri due metodi, le support vector machines potrebbero assegnare quasi tutte le osservazioni alla classe di maggioranza per due ragioni principali [14]:

- Gli eventi rari si trovano più lontani dal confine decisionale ideale rispetto alle istanze della classe prevalente;

- Il rapporto tra i vettori di supporto delle due classi è sbilanciato, perciò è più probabile che un'osservazione di test posizionata vicino al confine decisionale venga assegnata alla classe prevalente.

[12] King, G. e Zeng, L. (2001). Logistic regression in rare events data. Political Analysis.
[13] Hand, D.J. e Vinciotti, V. (2003). Choosing K for Two-Class Nearest Neighbour Classifiers with Unbalanced Classes.
[14] Wu, G. e Chang, E. (2003). Class-Boundary Alignment for Imbalanced Dataset Learning.

3.1 Trattamento dei dataset sbilanciati

Nel corso degli anni sono state proposte diverse soluzioni per risolvere il problema dello sbilanciamento. È possibile distinguere questi approcci in due categorie:

- Tecniche di Cost-Sensitive Learning, che attribuiscono un costo più elevato all'errata classificazione per la classe di minoranza utilizzando una matrice dei costi che ha la stessa struttura della matrice di confusione;

- Tecniche di campionamento che modificano il dataset originale in modo da ottenere una distribuzione bilanciata tra le due classi. Possono essere distinte in tecniche di oversampling nel caso in cui aggiungano istanze alla classe di minoranza e in tecniche di undersampling nel caso in cui eliminino delle osservazioni dalla classe prevalente.

Le tecniche di Cost-Sensitive Learning utilizzano delle penalità per le classificazioni errate attraverso una matrice di costo, utilizzata in fase di costruzione del modello. In genere le osservazioni vengono bilanciate attraverso l'attribuzione di un peso legato alla penalità della classe di appartenenza. Attraverso questa tecnica viene attribuito un costo superiore per l'errata classificazione di un'osservazione della classe rara rispetto al costo di errata classificazione di un'osservazione della classe prevalente. Uno degli svantaggi è dovuto all'impossibilità di implementare tale tecnica a tutti i modelli di apprendimento automatico. Le tecniche di campionamento non hanno questo limite, poiché agiscono in una fase precedente, di pre-elaborazione, perciò il modello viene costruito come se le osservazioni appartenessero a un insieme di dati bilanciato.

3.2 Random oversampling e random undersampling

Le tecniche di campionamento più comuni sono il random oversampling e il random undersampling. Il random oversampling è un metodo non euristico che agisce attraverso la replicazione casuale di osservazioni della classe di minoranza. Al contrario, il random undersampling elimina in maniera casuale un certo numero di osservazioni della classe di maggioranza. Il numero delle istanze da eliminare viene definito a priori a seconda del rapporto che si desidera ottenere tra le classi.

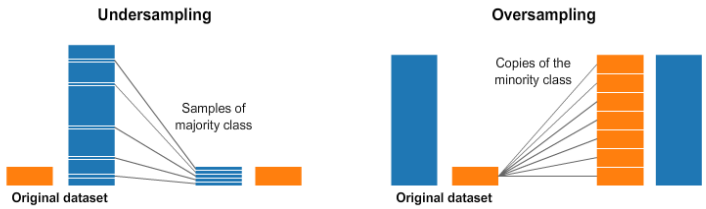


Figura 3.2 A sinistra: un esempio di Undersampling. A destra: un esempio di Oversampling. Fonte: https://www.pinterest.it/pin/514958538641697615/

Il meccanismo randomico utilizzato dai due metodi è simile ed entrambi possono generare lo stesso rapporto tra le classi di risposta. Tuttavia, ogni metodo presenta una serie di vantaggi e svantaggi differenti. Il random oversampling può aumentare il rischio di adattamento eccessivo poiché replica esattamente le osservazioni originali della classe di minoranza. In questo modo la regola decisionale utilizzata per la classificazione potrebbe sembrare accurata, ma in realtà lo è solamente poiché copre esempi duplicati. Inoltre, se il dataset utilizzato ha una numerosità elevata, i tempi di esecuzione degli algoritmi potrebbero diventare piuttosto lunghi. Al contrario, il random undersampling non aumenta il rischio di overfitting poiché non crea nuove osservazioni e può diminuire i tempi di computazione in quanto il numero delle osservazioni si riduce. Allo stesso tempo, l'esclusione delle osservazioni è la causa del problema principale del random undersampling. Infatti, ogni osservazione potrebbe contenere informazioni utili per discriminare correttamente le classi di risposta.[15] Nel corso degli anni sono stati sviluppati ulteriori metodi con l'obiettivo di superare i limiti di overfitting e perdita di informazioni utili. Di seguito viene descritto il funzionamento dei metodi SMOTE, ROSE e MWMOTE, mentre nel capitolo 4 sono stati applicati gli stessi metodi a un dataset che presenta un forte sbilanciamento tra le due classi di risposta.

[15] Kotsiantis, S., Kanellopoulos, D., Pintelas, P. (2006). Handling imbalanced dataset: A review.

3.3 SMOTE

Il metodo SMOTE (Synthetic Minority Over-sampling Technique) è una tecnica di oversampling introdotta nel 2002 in cui la classe di minoranza viene sovracampionata creando esempi artificiali, che vengono aggiunti al dataset originale prima della fase di apprendimento. Questo approccio si ispira a una tecnica utilizzata nel riconoscimento dei caratteri scritti a mano.[16] Le nuove osservazioni della classe di minoranza sono create utilizzando un criterio di similarità nello spazio dei predittori. Per ciascuna osservazione x_i della classe di minoranza vengono presi in considerazione i *K-nearest neighbors*, cioè le osservazioni della classe di minoranza la cui distanza euclidea da x_i ha il valore minimo lungo le p dimensioni dello spazio delle variabili indipendenti.

Il funzionamento di Smote può essere definito tramite la seguente funzione:

$$x_{new} = x_i + \lambda(x_{zi} - x_i) \tag{3.1}$$

In uno spazio di due dimensioni si supponga che (0.60, 3.55) sia un'osservazione della classe di minoranza x_i e (0.40, 3.25) il valore k più vicino x_{zi}, scelti casualmente. Si scelga un numero casuale λ compreso tra 0 e 1. La nuova osservazione x_{new} , calcolata con la (3.1) è mostrata nella figura 3.3.

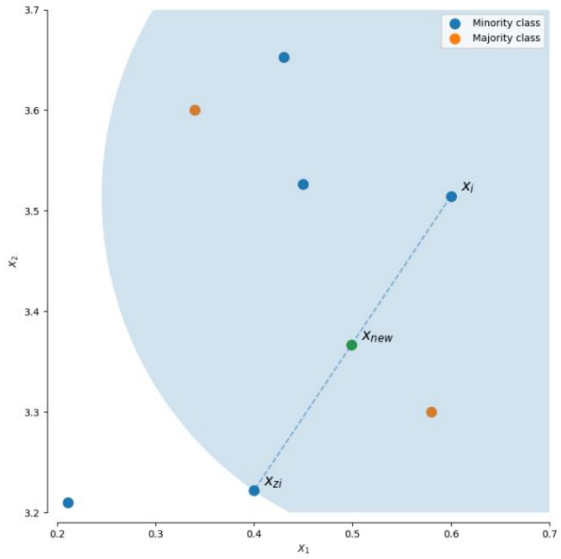


Figura 3.3 Un esempio di osservazione creata tramite il metodo SMOTE. Fonte: https://imbalanced-learn.org

[16] Chawla, N. V., et al. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, Vol. 16.

Il metodo SMOTE prende in considerazione cinque k-nearest neighbors e tra questi ne seleziona uno in maniera casuale, x_{zi}. Successivamente calcola la distanza euclidea tra x_{zi} e x_i e la moltiplica per un valore λ compreso tra 0 e 1 e scelto casualmente. Infine, questo valore viene sommato all'osservazione x_i. In questo modo viene creato il nuovo esempio artificiale, che si trova sulla linea che unisce x_i e x_{zi} nello spazio euclideo. I passaggi vengono ripetuti a seconda di quante nuove osservazioni si desidera creare.
I risultati ottenuti tramite questo metodo hanno mostrato un miglioramento delle prestazioni al crescere della percentuale di bilanciamento, fino al 200%. L'utilizzo di percentuali più elevate non ha evidenziato ulteriori miglioramenti, fino a peggiorare le prestazioni rispetto al dataset originale nei casi in cui è stata utilizzata una percentuale del 500%.
I vantaggi principali dall'utilizzo di questo metodo rispetto sono:

- Il rischio di overfitting rispetto al random oversampling è minore poiché le osservazioni create non sono la replica di altre osservazioni;

- Non vi è una perdita di informazioni dovuta all'esclusione di alcune osservazioni della classe di maggioranza.

Tra gli svantaggi, nel caso in cui si utilizzi una percentuale di bilanciamento elevata, vi è il rischio di *overgeneralization*. SMOTE non prende in considerazione le osservazioni della classe di maggioranza durante il suo processo di creazione, perciò potrebbe verificarsi una situazione in cui le osservazioni sintetiche invadano lo spazio dell'altra classe, rendendo difficile la distinzione durante la classificazione.

3.4 ROSE

Nel 2012 Menardi e Torelli hanno proposto un nuovo metodo di campionamento, chiamato Random Over Sampling Examples (ROSE). Anche questo approccio è basato sulla generazione di nuove osservazioni artificiali, ma utilizza la stima di una funzione di smoothing kernel.[17] Si supponga che $n_j < n$ sia la dimensione della classe Y_j con $j = 0, 1$. Il metodo ROSE genera le nuove istanze attraverso le seguenti fasi:

- Seleziona casualmente una delle classi Y_j, a cui viene assegnata la stessa probabilità del 50%;
- Estrae casualmente un'osservazione x_i, selezionata dal sottoinsieme delle unità appartenenti alla classe di risposta scelta casualmente nella fase precedente, con probabilità $p_i = {}^{1}/_{n_j}$
- Stima una distribuzione dell'intorno di x_i attraverso una funzione kernel K_{H_j}[18] con media x_i e matrice di covarianza H_j:

$$\hat{f}(x|y = Y_j) = \sum_i^{n_j} p_i Pr(x|x_i) = \sum_i^{n_j} \frac{1}{n_j} Pr(x|x_i) = \sum_i^{n_j} \frac{1}{n_j} K_{H_j}(x - x_i) \quad (3.2)$$

- Estrae casualmente la nuova osservazione sintetica da una funzione di probabilità proporzionale a K_{H_j}.

In sostanza, viene scelta un'osservazione appartenente a una delle due classi, assegnando la stessa probabilità a Y_1 e Y_2, e generata una nuova istanza nel suo intorno, dove l'ampiezza dell'intorno è determinata da H_j. Anche in questo caso la ripetizione delle fasi appena descritte consente la creazione di un nuovo training set T_m^*, in cui il numero di osservazioni tra le due classi di risposta è approssimativamente uguale. La dimensione m del training set può essere impostata in base alla numerosità del training set originale oppure scelta in qualsiasi altro modo. SMOTE non chiarisce le motivazioni alla base del processo di generazione delle istanze artificiali e giustifica la scelta del metodo solamente tramite ragioni euristiche. Al contrario, ROSE giustifica il processo di bilanciamento attraverso basi

[17] Menardi, G., Torelli, N., (2012). Training and assessing classification rules with unbalanced data.
[18] Mauriello F. (2014), Tecniche di ricampionamento per dataset con classi di risposta sbilanciate. Una proposta metodologica per dataset con predittori di natura numerica e categorica

teoriche, supportate dalle proprietà dei metodi dei kernel. Inoltre, ROSE non si focalizza solamente sulla classe di minoranza, ma combina tecniche di oversampling e undersampling per ottenere un dataset bilanciato, come mostrato nella figura 3.4.

Figura 3.4 Un esempio del funzionamento dei metodi di oversampling, undersampling, SMOTE e ROSE.
Fonte: Tantithamthavorn, C. et al. (2020), The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models.

3.5 MWMOTE

Il metodo MWMOTE è stato introdotto nel 2014, al fine di superare alcuni limiti delle tecniche di oversampling sintetico.[19] Il problema, già menzionato tra gli svantaggi di SMOTE, si verifica poiché questi approcci utilizzano alla cieca tutti i k-nearest neighbors senza considerare la loro distanza e posizione. Tale problema è amplificato quando le osservazioni della classe di minoranza sono raggruppate in cluster di piccole dimensioni, come mostrato nella figura 3.4. Infatti, se x_i è un'osservazione del cluster L2, x_{zi} potrebbe essere scelta casualmente tra le osservazioni del cluster L1. Si può notare che l'osservazione creata, in questo caso R, si sovrappone alla classe di maggioranza e tale situazione potrebbe rendere difficile il processo di apprendimento.

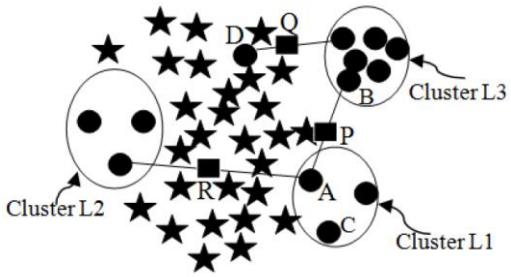


Figura 3.4 Un esempio dei possibili problemi nell'utilizzo di un metodo sintetico. Fonte: Barua S. et. Al (2014), MWMOTE- Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning

Inoltre, se l'osservazione viene selezionata da un cluster in cui le osservazioni sono particolarmente vicine, come il punto B del cluster L3 della figura 3.4, per un determinato valore di k ≤ 5 le nuove osservazioni potrebbero essere molto simili a quelle esistenti. In questo modo si avrebbero istanze duplicate, che non aggiungono nuove informazioni alla classe di minoranza. L'obiettivo di MWMOTE è duplice: migliorare la selezione del campione x_i e migliorare il processo per la creazione delle nuove istanze artificiali. È possibile identificare tre fasi principali:

- Identifica le osservazioni della classe di minoranza più importanti, S_{min}, e costruisce l'insieme di tali osservazioni, S_{imin};
- Assegna un peso, S_w, a ogni osservazione di S_{imin}, in base alla sua importanza;
- Genera le nuove osservazioni da S_{imin}, in base agli S_w e produce un nuovo dataset S_{omin}, aggiungendo le nuove osservazioni artificiali a S_{min}.

[19] Barua S. et. Al (2014), MWMOTE- Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning.

Le osservazioni più importanti sono quelle che il modello potrebbe non apprendere facilmente, poiché si trovano vicino al confine decisionale o appartengono ai cluster di piccole dimensioni. Al fine di selezionare tali osservazioni, si rimuovono dall'insieme S_{min} tutti gli x_i della classe di minoranza per il quale, definito il valore di k, i k-nearest neighbors appartengono alla classe di maggioranza. In questo modo vengono eliminate tutte le osservazioni che potrebbero causare una sovrapposizione, come mostrato nell'esempio precedente. L'insieme creato dopo la rimozione di tali osservazioni è definito $S_{\min f}$. Successivamente, per ogni campione dell'insieme $S_{\min f}$ vengono identificati tutti i k-nearest neighbors della classe di maggioranza, che presumibilmente si trovano vicino al confine decisionale. La somma di queste osservazioni costituisce S_{bmaj}. Infine, dall'ultimo insieme creato si calcolano i k-nearest neighbors della classe di minoranza, per formare S_{imin}. Nella figura 3.5 è mostrato un esempio del processo appena descritto.

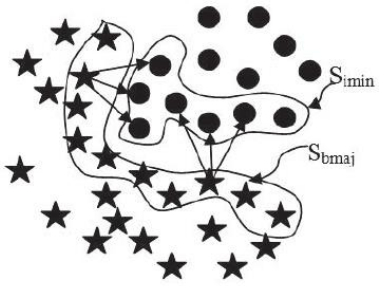


Figura 3.5 Un esempio della prima fase di MWOTE. Si identificano gli insiemi S_{imin} e S_{bmaj}. Fonte: Barua S. et. Al (2014), MWMOTE- Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning

Dopo aver stabilito l'insieme delle osservazioni S_{imin}, è necessario considerare che ciascuna di esse potrebbe avere un'importanza diversa. Per tale ragione, MWMOTE assegna un peso S_w che influenza il numero delle istanze x_{new} generate da un'osservazione x_i, sulla base delle seguenti considerazioni:

- Le osservazioni più vicine al confine decisionale contengono più informazioni, perciò dovrebbero avere un valore di S_w maggiore;
- Le osservazioni che si trovano in cluster con bassa densità sono più importanti rispetto a quelle di un cluster con una densità elevata, poiché quest'ultimo contiene più informazioni. Per ridurre tale squilibrio viene assegnato un S_w maggiore alle istanze del primo cluster;
- Le osservazioni vicine a un cluster della classe di maggioranza con una densità elevata sono più importanti rispetto a quelle vicine a un cluster della classe di

maggioranza con una densità bassa, poiché tale vicinanza potrebbe complicare il processo di apprendimento.

Il calcolo di S_w è determinato dalla seguente formula:

$$S_w(x_i) = \sum_{z_i \in S_{bmaj}} I_w(z_i, x_i) \tag{3.3}$$

In cui $I_w(z_i, x_i)$ è il prodotto di un fattore di vicinanza e di un fattore di densità tra le osservazioni dell'insieme S_{bmaj} e quelle dell'insieme S_{imin}. Infine, MWOTE genera le nuove istanze artificiali tramite la (3.1). Nonostante l'equazione utilizzata sia la stessa di SMOTE, il vantaggio di tale metodo deriva dalla selezione di x_{zi}. Infatti, tramite un approccio basato sul clustering gerarchico modificato, si evita il problema di sovrapposizione rappresentato nella figura 3.4, poiché x_{zi} è selezionato dal medesimo cluster di x_i. Un altro problema evidenziato in precedenza è il rischio che le osservazioni artificiali siano molto simili a quelle esistenti. Tale rischio è mitigato dal peso S_w inferiore assegnato alle osservazioni incluse nel cluster con una densità elevata. La figura 3.5 mostra la differenza tra MWMOTE e i metodi che creano le istanze scegliendo x_{zi} casualmente tra i k-nearest neighbors. È evidente che nell'immagine a sinistra vi è il rischio di sovrapposizione con le osservazioni della classe di maggioranza.

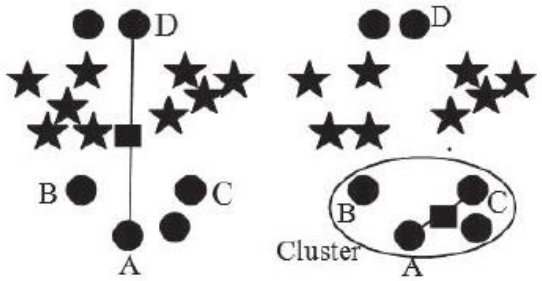


Figura 3.5 La figura illustra la differenza tra un metodo di oversampling sintetico (a sinistra) e MWMOTE (a destra). Fonte: Barua S. et. Al (2014), MWMOTE- Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning

4. Il caso Google Store

4.1 Descrizione del dataset

Il dataset analizzato è relativo alle visite e le relative transazioni effettuate nel Google Store tra il 01-08-2016 e il 01-08-2017. Sono presenti 903.653 osservazioni, ma per motivi computazionali è stato estratto un campione casuale di 25.000 osservazioni. Al termine della fase di cleaning, che verrà descritta in seguito, il dataset è costituito dalle seguenti 23 variabili:

- transactionRevenue: la variabile target che indica il valore in dollari delle singole transazioni nel Gstore;
- fullVisitorId: rappresenta un ID univoco per ogni utente del Gstore, noto anche come ID cliente;
- visitId: è un numero che identifica in maniera univoca la visita di un utente. Se due utenti visitassero l'e-commerce nello stesso momento, considerato anche il fuso orario, avrebbero lo stesso visitId. Pertanto, per garantire l'univocità è necessario utilizzare una combinazione di fullVisitorId e visitId;
- sessionId: è generato dalla combinazione delle variabili precedenti e permette di identificare in maniera univoca sia la sessione che l'utente del Gstore;
- Date: indica il giorno, il mese e l'anno della sessione;
- VisitStartTime: fornisce le stesse informazioni della variabile Date, aggiungendo ulteriori dettagli in termini di ore, minuti e secondi. In alcuni casi la data della sessione potrebbe differire a causa del fuso orario differente per ogni singola timezone.
- channelGrouping: indica la categoria del canale utilizzato dall'utente per visitare il Gstore. Sono stati individuate 8 categorie differenti: Affiliates, Direct, Display, Paid Search, Referral, Social e Altri;
- visitNumber: rappresenta il numero della sessione per l'utente. Se è la prima sessione, questo numero è impostato su 1;
- Visits: indica se nella sessione ci sono stati eventi di interazione;
- Source: contiene informazioni sulla sorgente di traffico che ha generato la sessione. Il traffico di un sito web deve provenire da una determinata fonte, che si tratti di persone che visitano il sito dai motori di ricerca, o da un sito di social media o da qualche altro sito web. Quando non proviene da un sito Web o non ci sono dati sul sito Web originale,

la fonte è nota come Direct. Durante la procedura di cleaning le osservazioni sono state raggruppate in 4 categorie principali: Google, Youtube, Direct e Altri;

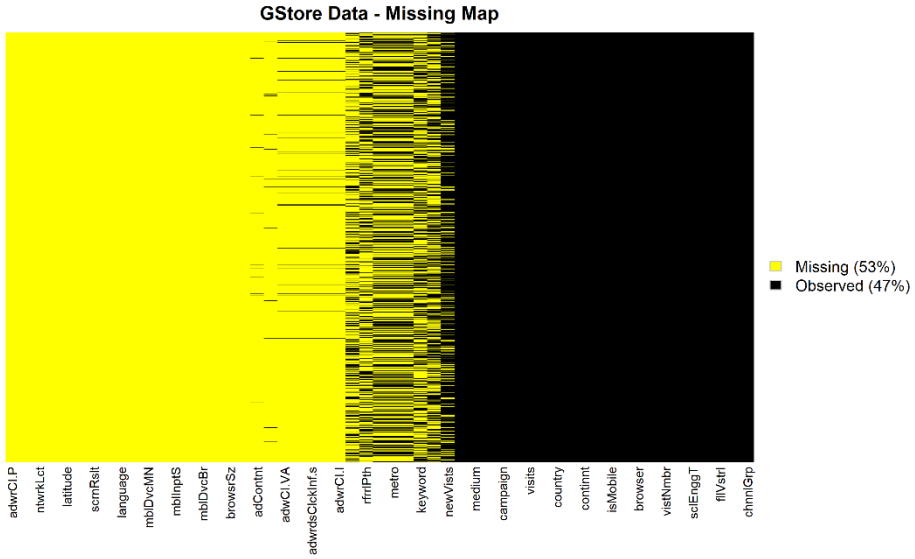
- isTrueDirect: assume il valore True se la sorgente della sessione è diretta, ovvero se l'utente ha digitato il nome dell'URL del tuo sito web nel browser o è arrivato al sito tramite un segnalibro, altrimenti sarà False;
- Medium: rappresenta il mezzo della sorgente di traffico. Può essere "organico", "cpc", "referral", "cpm", "affiliate" e "unknown";
- Pageviews: indica il numero totale di pagine visualizzate in una singola sessione;
- Hits: indica il numero delle interazioni dell'utente durante la sessione. Per ogni pagina visitata possono essere generate più interazioni, in quanto oltre il testo sono presenti ulteriori assets, ad esempio immagini o video. Per ognuno di questi assets viene inviata una richiesta al server che sarà conteggiata come interazione dell'utente;
- Variabili di geolocalizzazione: Continent, subContinent e Country indicano la posizione geografica dell'utente;
- OperatingSystem: fornisce informazioni sul sistema operativo del dispositivo utilizzato per effettuare l'accesso. Le osservazioni sono state raggruppate in 8 categorie differenti: Android, Chrome OS, iOS, Linux, Macintosh, Windows, Windows Phone e Other OS, che include tutti i sistemi operativi poco diffusi;
- Browser: indica il browser web utilizzato dall'utente. Anche in questo caso sono state individuate 10 categorie: Android Webview, Chrome, Edge, Firefox, Internet Explorer, Opera, Opera Mini, Safari, Safari (in-app) e Other, che include i browser web poco diffusi;
- DeviceCategory: fornisce informazioni sulla categoria del dispositivo. Può essere desktop, mobile o tablet;
- isMobile: il valore della variabile sarà True se il dispositivo utilizzato è uno smartphone o un tablet, False se il dispositivo è desktop.

Attraverso la variabile fullVisitorID è possibile verificare il numero di utenti, che differisce dal numero di osservazioni poiché un singolo ID potrebbe aver visitato il Gstore più di una volta nell'arco temporale analizzato. Infatti, tra le 20.000 osservazioni del training set sono presenti 19625 utenti che hanno visitato il Gstore almeno una volta durante l'arco temporale analizzato.

4.2 Cleaning dei dati

Prima di procedere con l'analisi del dataset, è stato eseguito un processo di *data cleaning* che ha permesso di verificare la correttezza dei dati, la presenza di dati mancanti e di apportare ulteriori modifiche che potessero agevolare l'analisi stessa**.**

Attraverso la seguente rappresentazione grafica è possibile individuare la presenza di dati mancanti per ogni variabile del dataset. Nell'asse delle ascisse sono indicate le variabili e nell' asse delle ordinate le osservazioni. Le linee gialle del grafico rappresentano un dato mancante, invece, le linee nere indicano la presenza di un determinato valore.



In primo luogo, sono state escluse le variabili che presentavano esclusivamente dati mancanti, poiché non avrebbero fornito alcuna informazione utile ai fini dell'analisi. Per le medesime ragioni, sono state eliminate le variabili con una percentuale di dati mancanti elevata, qualora non fosse possibile la sostituzione con altri valori. In seguito, sono state analizzate le variabili per le quali è stato possibile attribuire, attraverso la sostituzione, un valore significativo ai dati mancanti:

- transactionRevenue: per la variabile che indica l'importo degli acquisti in una determinata sessione i valori mancanti sono stati sostituiti dal numero 0, per indicare un ricavo pari a 0$ qualora l'utente non avesse effettuato nessun acquisto durante la sessione. Tale variabile è stata dicotomizzata e presenta il valore 1 nel caso in cui la sessione si sia conclusa con un acquisto, 0 in caso contrario.

- *newVisits:* i valori mancanti sono stati sostituiti dal valore 0, poiché si tratta di una variabile dummy che presenta il valore 1 per indicare la presenza dell'attributo e un valore *NA* in caso di assenza;
- *isTrueDirect*: in corrispondenza delle linee nere il dataset indicava il valore *TRUE,* perciò i dati mancanti sono stati sostituiti con il valore *FALSE* per indicare l'assenza dell'attributo;
- *pageviews:* indica il numero di pagine visitate dall'utente in una determinata sessione, perciò il valore minimo non può essere inferiore a una singola pagina. Dopo aver verificato che in assenza di dati per le pagine visitate, il numero di interazioni (*hits)* è pari a 1, si è optato per la sostituzione dei dati mancanti con il valore minimo 1.

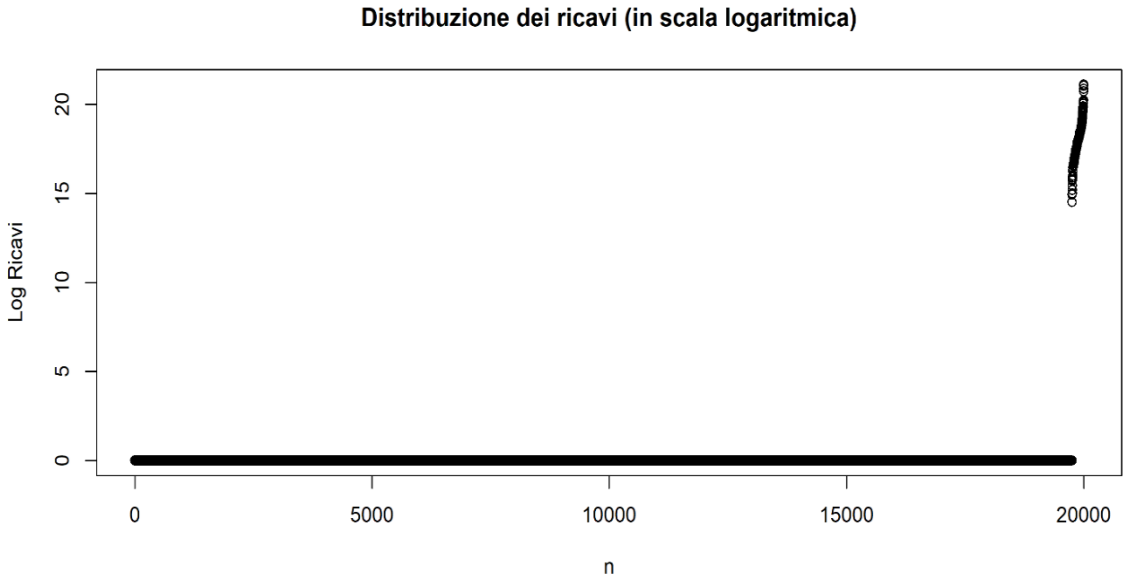
Al fine di agevolare l'analisi esplorativa e l'interpretazione delle rappresentazioni grafiche è stato ridotto il numero di modalità di alcune variabili, raggruppando le osservazioni che presentavano alcune caratteristiche comuni:

- per la variabile *source* il 90% delle osservazioni è stata reindirizzata da Url riconducibili a *Google, Youtube* o da fonti non meglio identificate (*Direct).* Per tale ragione sono state individuate 4 modalità: *Google, Youtube, Direct* e Other, per il 10% delle osservazioni residue;
- Il 2% degli utenti ha utilizzato alcuni browser con frequenze inferiori all'1%. Per tali osservazioni è stata creata un'unica modalità: *Altri.*
- Il 2% degli utenti ha utilizzato alcuni sistemi operativi, che presentavano frequenze relative piuttosto basse e inferiori all'1%. Tali osservazioni sono state raggruppate in un'unica modalità: *Altri OS.*

Il dataset è stato diviso in train e test sulla base della variabile "visitStartTime", che indica la data e l'orario di inizio della sessione. Nel train sono state incluse le osservazioni dal 01-08-2016 al 12-05-2017 e nel test le osservazioni successive, fino al 01-08-2017.

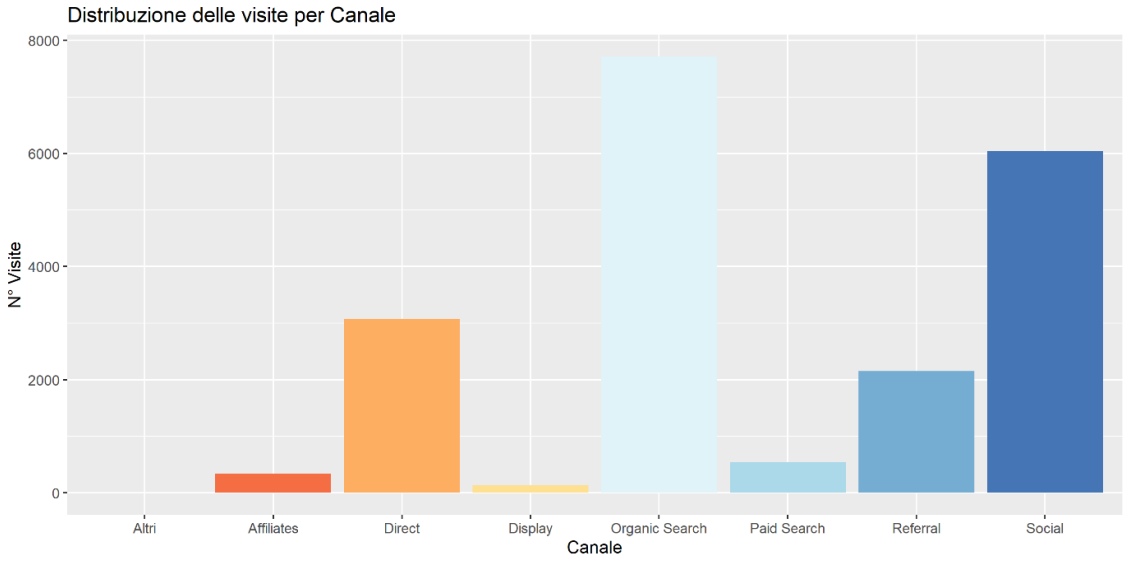
4.3 Analisi esplorativa

L'obiettivo dell'analisi esplorativa è verificare la distribuzione dei ricavi e delle altre variabili in termini di visite e ricavi generati. Dalla distribuzione logaritmica dei ricavi è evidente che gran parte delle visite non si conclude con un acquisto. Infatti, nel campione di 20.000 osservazioni, solamente 243 visite hanno generato un ricavo.



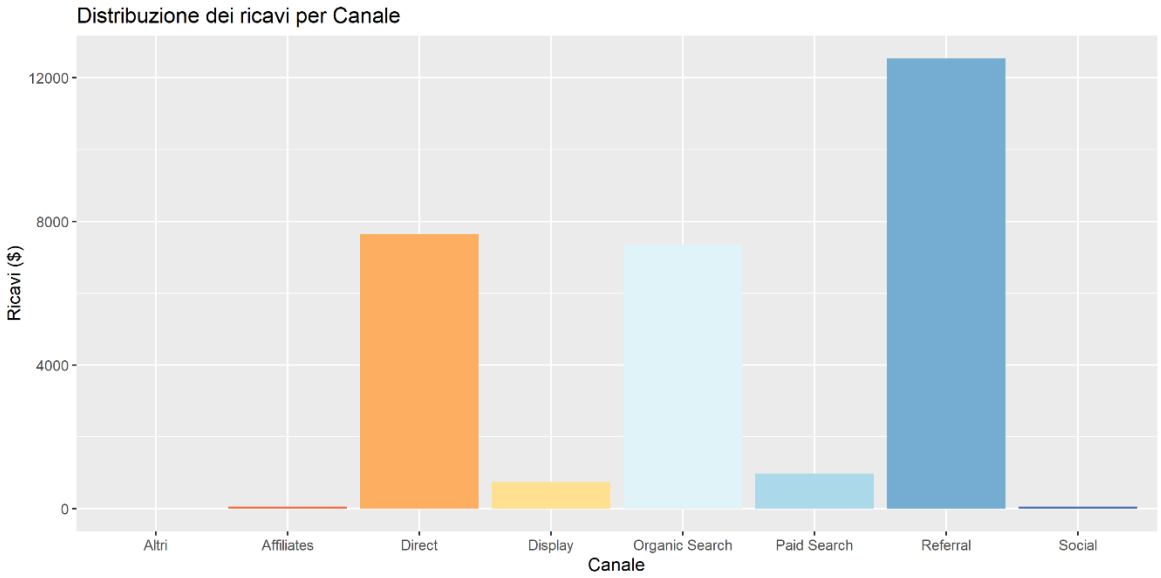
4.3.1 Channel Grouping

Il primo grafico mostra le frequenze percentuali in termini di numero di visite per la variabile "Channel Grouping". Il canale maggiormente utilizzato (~40%) è la ricerca organica dell'e-commerce, seguito dai Social Network (~30%), dal canale Direct (~15%) e dai Referral (~12%).



Per ricerca organica si intendono gli accessi effettuati da una query sui motori di ricerca (Google, Bing, Yahoo, ecc.), esclusi gli annunci sponsorizzati (Paid Search).

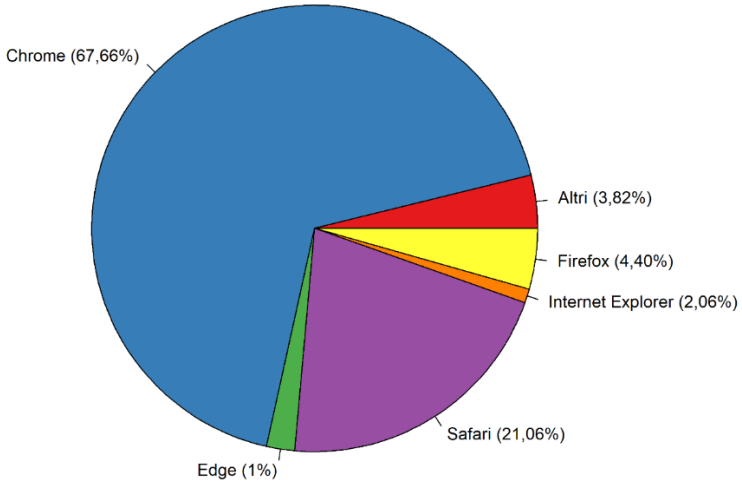
Il canale Direct è di difficile interpretazione poiché non include solamente gli accessi diretti degli utenti che digitano l'URL nella barra degli indirizzi, ma si ottiene per sottrazione. Se la fonte del traffico non è attribuibile né ad una campagna, né ad una sorgente specifica viene catalogato da Google Analytics come "Direct". Si rilevano particolarmente validi i Social Network e i Referral, che possono provenire da una serie di sorgenti, tra cui post di Google Gruppi o pagine statiche su siti Google correlati. Inoltre, è evidente il contributo piuttosto basso in termini di visite degli annunci sponsorizzati (Paid Search), come Google Adwords.



Nonostante il numero di visite generate dalla ricerca organica sia particolarmente elevato, dal grafico è visibile che non sia il più remunerativo per l'e-commerce. Il contributo dei Referral è particolarmente elevato, con solo il 12% delle visite totali contribuisce in maniera sostanziale ai ricavi del Gstore, ma è sorprendente il contributo del canale Display, costituito dai banner pubblicitari a pagamento, con un numero di visite inferiore al 5% è il secondo canale per ricavi generati.

4.3.2 Browser

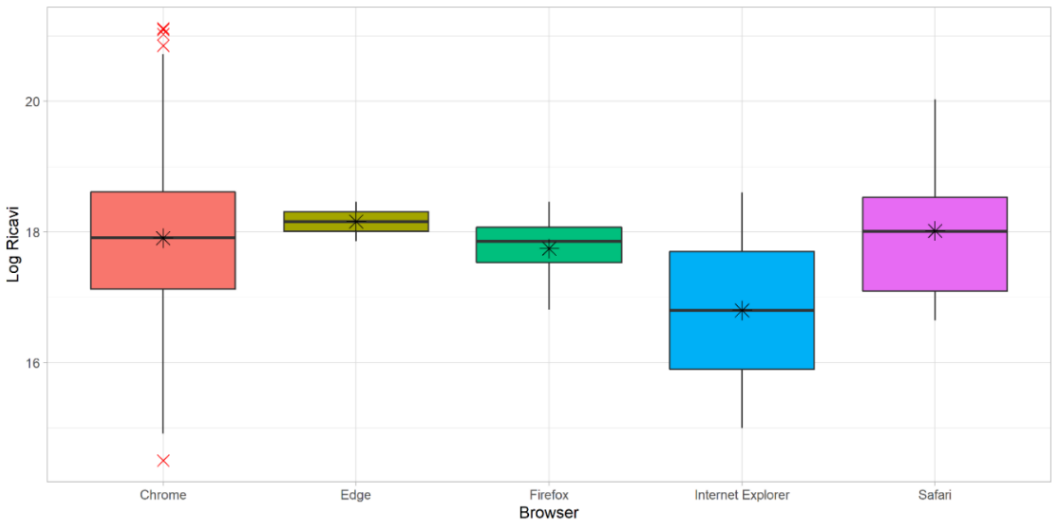
Dall'analisi della distribuzione delle visite a seconda del browser utilizzato è evidente il maggior impiego di Google Chrome (67,66%) rispetto agli altri browser:
Safari è utilizzato dal 21,2% degli utenti, invece, Firefox, Internet Explorer ed Edge contribuiscono al 7,46% delle visite.



La percentuale di diffusione di Google Chrome può essere in linea con le aspettative, per un probabile maggior interesse nei prodotti del Gstore con il brand del browser utilizzato. Il numero delle visite da Safari è oltre le aspettative, poiché si tratta di un browser nativo dei prodotti Apple, concorrente di Google per diversi prodotti e mercati.

Browser	Ricavi
Chrome	27123,71 $
Edge	160,68 $
Firefox	237,87 $
Internet Explorer	123,25 $
Safari	1714,45 $

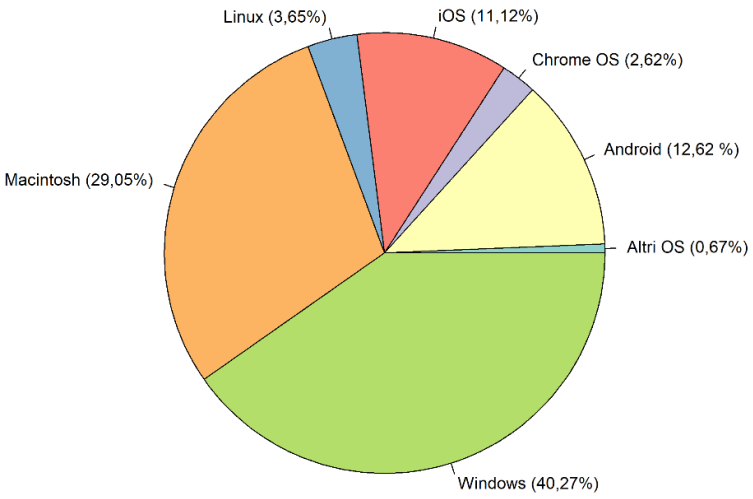
Dalla tabella dei ricavi generati da ciascun browser si nota che Chrome è il browser che contribuisce ai ricavi del Gstore in misura superiore (~27.000 $), con un vantaggio dato dal numero di visite più elevato rispetto agli altri browser. In questo caso l'ordine nei ricavi rispetta le percentuali delle visite analizzate in precedenza.



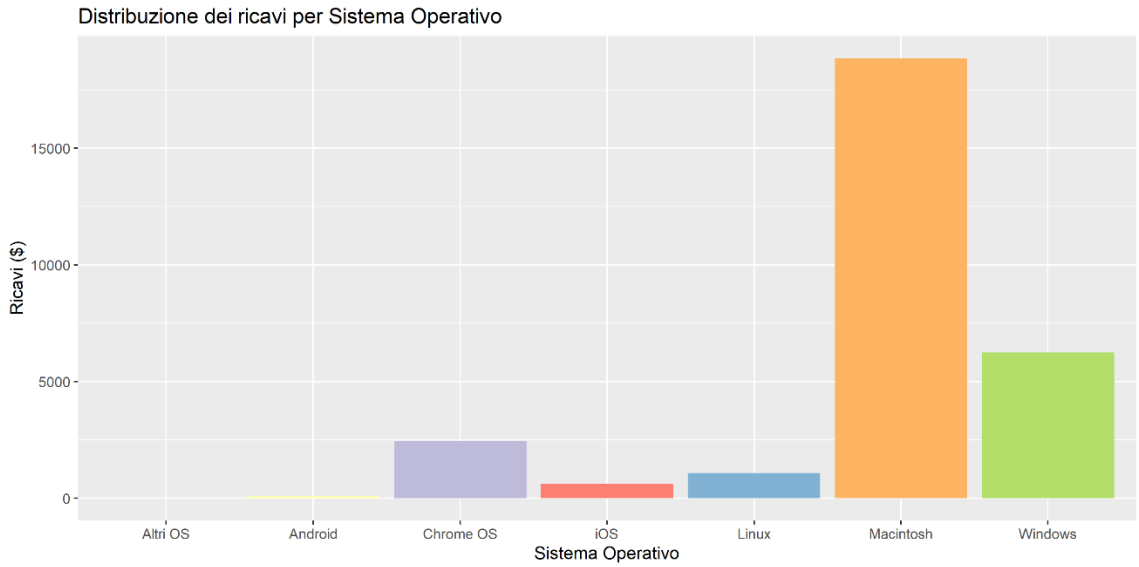
Nel boxplot sono rappresentate le osservazioni degli utenti che hanno effettuato un acquisto nel Gstore, suddivise per il browser utilizzato. Per agevolare la rappresentazione grafica i ricavi sono stati trasformati in logaritmo naturale. È possibile notare la posizione della mediana simile per tutti i browser, tranne Internet Explorer. Per ciascun browser il valore medio si trova in corrispondenza della mediana, ad eccezione di Firefox, in cui la media ha un valore inferiore. Edge e Firefox hanno una dispersione piuttosto bassa, invece, Chrome ha una dispersione elevata e alcuni outliers.

4.3.3 Operating System

Dall'analisi del sistema operativo più utilizzato si nota la prevalenza di Windows (40,27%) e Macintosh (29,05%). Il restante 30,68% è distribuito tra Android (12,62%) e iOS (11,12%) ed una percentuale residua tra Linux, Chrome Os e altri sistemi operativi.

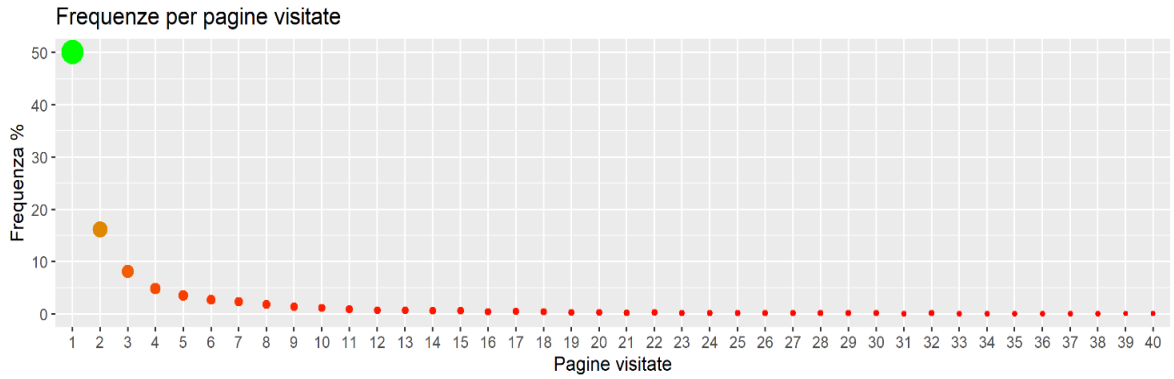


Android e Chrome OS, entrambi progettati e sviluppati da Google, occupano una posizione marginale. È inoltre evidente il minor utilizzo di dispositivi mobili (smartphone e tablet): oltre il 70% delle visite è stato effettuato da sistemi operativi installati su desktop e notebook.

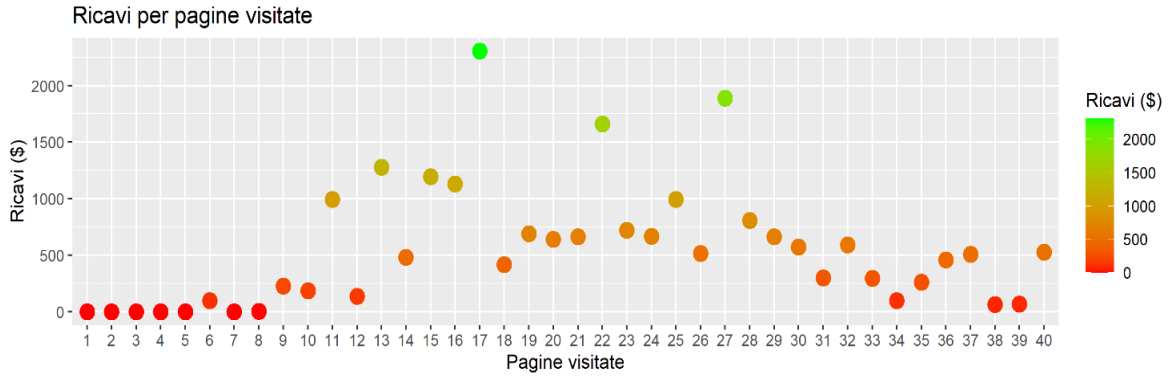


Il maggior volume dei ricavi è prodotto dalle visite dai dispositivi Macintosh e Windows, che confermano la preminenza evidenziata nel grafico relativo alle visite. Nonostante il numero di visite inferiore, dagli utenti Macintosh sono stati generati quasi 13.000$ in più rispetto ai ricavi generati dagli utenti Windows. I ricavi generati da Chrome OS, che ha una frequenza di utilizzo inferiore al 5%, sono inferiori solamente a Windows e Macintosh: probabilmente questi utenti sono particolarmente interessati al marchio Google e ai suoi prodotti. Inoltre, prendendo in considerazione i dispositivi mobili, i risultati evidenziano maggiori ricavi generati dai dispositivi Apple (iOS) rispetto ai dispositivi con un software Google (Android), nonostante il numero di visite leggermente più basso. Questo risultato è contro le aspettative, che vedrebbero un maggior volume di spesa generato da coloro che utilizzano un software Google.

4.3.4 Pageviews



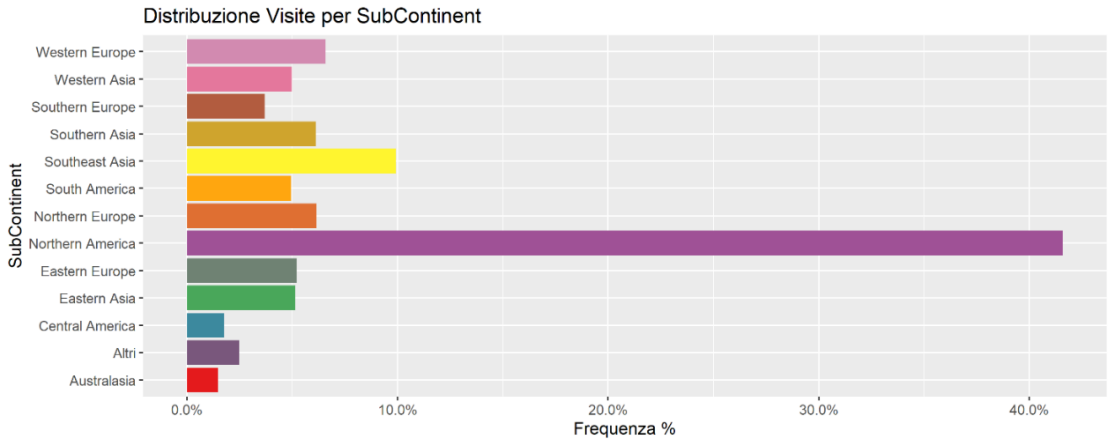
La maggioranza dei visitatori ha visitato una sola pagina e generato una sola interazione. La distribuzione per entrambe le variabili è decrescente, infatti, al crescere delle pagine e delle interazioni diminuisce la percentuale delle visite. Dal grafico è possibile calcolare una frequenza di rimbalzo pari al 50%: 10.000 utenti del campione osservato visitano una sola pagina e non proseguono la loro navigazione all'interno del Gstore.



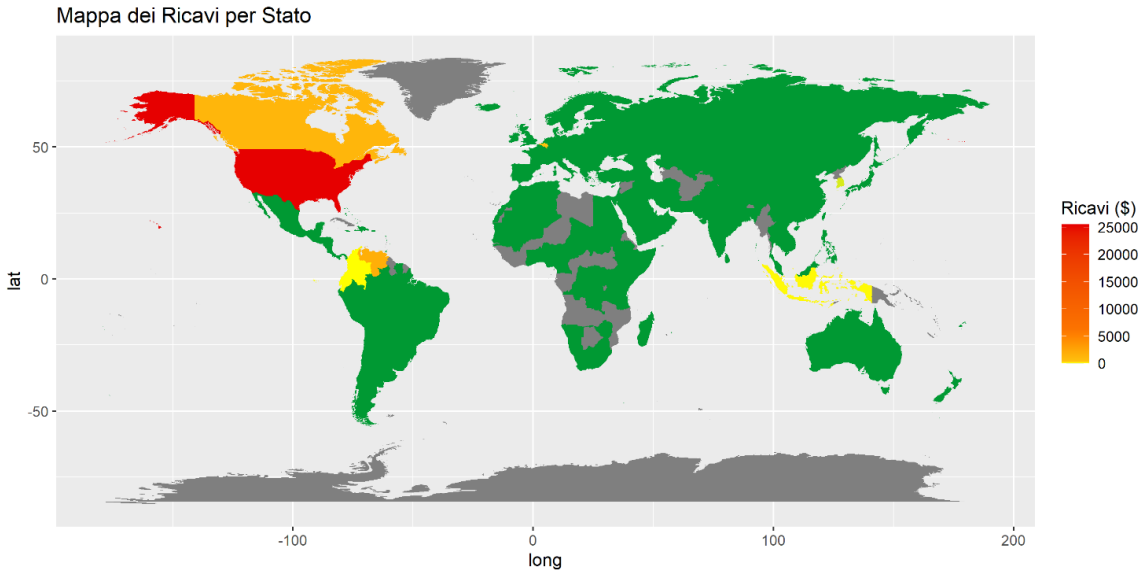
Per quanto concerne la distribuzione dei ricavi generati si può notare che ad un numero molto basso di visite corrispondono ricavi nulli o molto bassi. A partire da 8 pagine visitate i ricavi aumentano, ma non è possibile identificare una chiara tendenza o un modello.

4.3.5 Continent e SubContinent

I grafici di seguito mostrano la distribuzione geografica degli utenti del Gstore: oltre il 40% delle visite proviene dall'America del Nord, seguita dal sud-est asiatico (~10%). Le zone europee, suddivise in Ovest, Est, Nord e Sud, raggiungono una percentuale di circa il 5/6% ciascuna e del 21,6% totale. Nella categoria Australasia sono incluse l'Australia e la Nuova Zelanda che raggiungono una percentuale piuttosto bassa, il 2%. Inoltre, le regioni africane sono state raggruppate nella categoria Altri, poiché le singole zone del continente africano non raggiungevano l'1% delle visite totali.

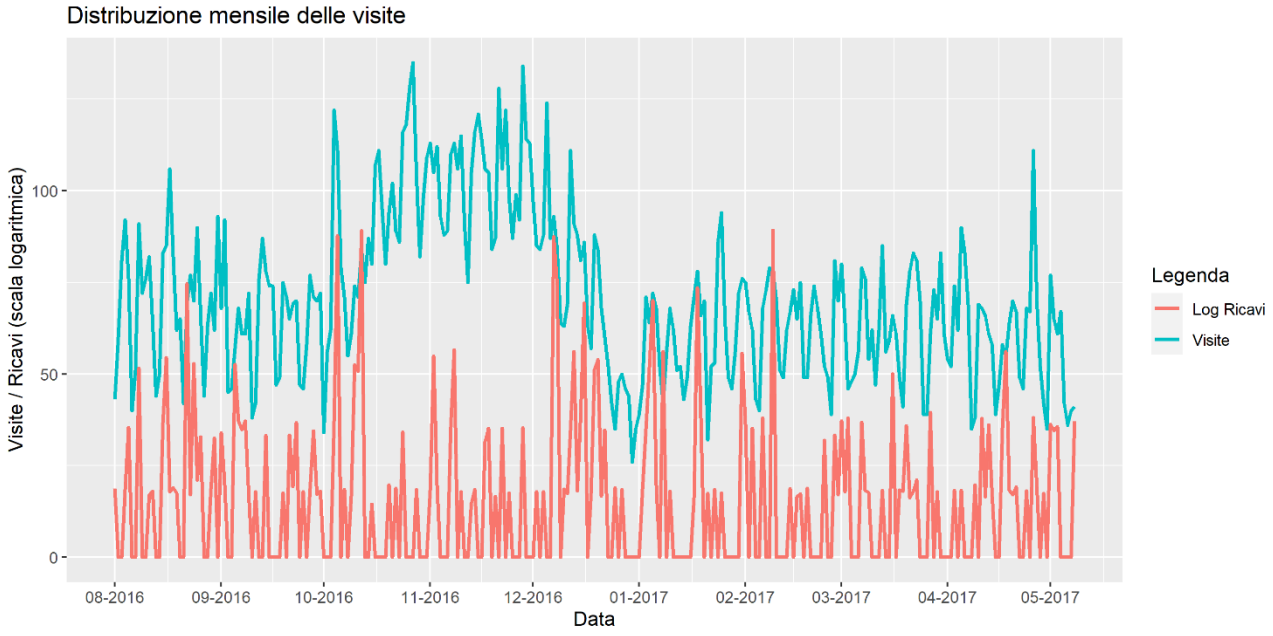


La distribuzione dei ricavi evidenzia una distribuzione degli acquisti in pochi Stati. La maggior parte dei ricavi è generata dagli Stati Uniti, seguiti dal Canada e dal Venezuela. Oltra al Venezuela, nel Sud America sono stati effettuati acquisti in Ecuador e Colombia. In Europa l'unica nazione in cui sono stati effettuati acquisti è il Belgio. In Asia sono stati effettuati acquisti solamente in Corea del Sud e Indonesia.

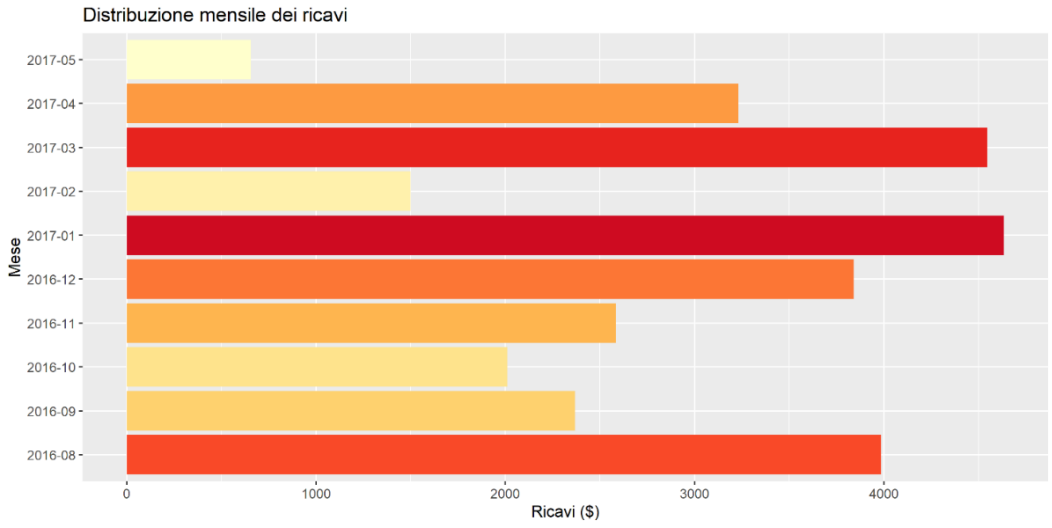


4.3.6 Distribuzione mensile

Nella distribuzione mensile delle visite si riscontra un aumento da ottobre a novembre 2016; in particolare a novembre si riscontra una percentuale superiore al 15% delle visite totali. A dicembre 2016 le visite calano fino a raggiungere le stesse frequenze osservate ad agosto e settembre 2016.

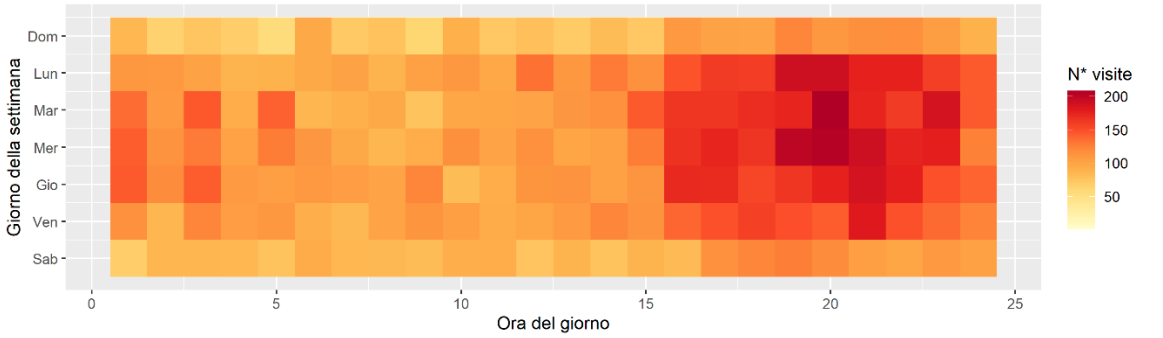


Per quanto concerne l'andamento dei ricavi, dal grafico precedente non è possibile evidenziare una tendenza chiara e costante né in aumento né in diminuzione, poiché sono presenti alcuni giorni nei quali gli utenti non hanno compiuto alcun acquisto. Invece, nel grafico successivo è possibile notare che durante il mese di novembre sono stati generati ricavi piuttosto bassi, nonostante il numero di visite elevato. Il mese con i ricavi più elevati è gennaio 2017, seguito da marzo 2017 e agosto 2018.

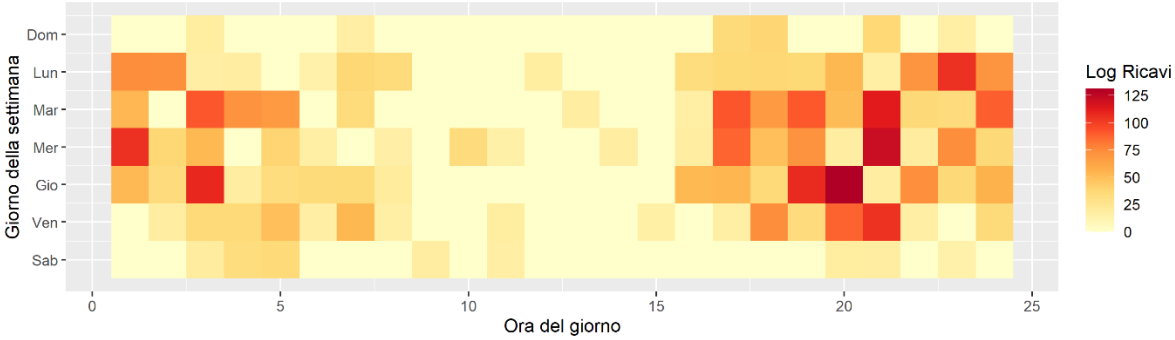


4.3.7 Distribuzione settimanale

Le osservazioni sono state raggruppate per giorno della settimana, evidenziando un numero di visite più basso durante il weekend. Infatti, le visite iniziano a decrescere il venerdì, raggiungendo le percentuali più basse il sabato e la domenica. Inoltre, è possibile notare un aumento delle visite dalle 16 alle 23 dell'intera settimana. Le frequenze più elevate sono state raggiunte tra le 19 e le 21 di martedì e mercoledì.



Per quanto concerne i ricavi è possibile confermare quanto evidenziato per le visite: durante il weekend i ricavi generati sono piuttosto bassi. Per ciascun giorno della settimana dalle 06.00 alle 16.00 gli utenti non hanno effettuato alcun acquisto oppure acquisti di importi bassi. Il giovedì, in particolare tra le 19 e le 20, è il giorno della settimana nel quale l'importo dei ricavi generati è più elevato.



4.4 Analisi predittiva

In questo paragrafo vengono presentati i risultati ottenuti dall'applicazione di tre modelli di classificazione: regressione logistica, analisi discriminante lineare e support vector machines. Inoltre, viene fatto un confronto tra le diverse tecniche di gestione dello sbilanciamento delle classi di risposta.

4.4.1 Bilanciamento del dataset

La classe di risposta è molto sbilanciata, infatti, se si considera l'intero dataset, la classe non acquisto ha una frequenza del 98.7%, mentre la classe acquisto ha una frequenza del 1.3%. A causa della distribuzione della variabile di risposta eccessivamente sbilanciata, il processo di apprendimento potrebbe essere distorto, perché il modello tenderebbe a focalizzarsi sulla classe prevalente. Per tale ragione sono state applicate tre tecniche di bilanciamento: Mwmote, Rose, Smote. Tali metodi hanno riguardato solamente il training set, cioè le istanze sono state create esclusivamente in fase di costruzione dei modelli, mentre per la validazione è stato utilizzato il test set originale. Un limite, che accomuna tutte le tecniche appena citate, è rappresentato dalla impossibilità di trattare dataset con classi sbilanciate in cui vi sia la presenza di variabili qualitative. Per tale ragione si è resa necessaria l'esclusione di tali variabili prima di applicare le diverse tecniche di bilanciamento.

	Training set		Test set	
Bilanciamento	Acquisto	Non acquisto	Acquisto	Non acquisto
Originale	19739	261	4938	62
Numerico	19739	261	4938	62
Mwmote	19739	19739	4938	62
Rose	9997	10003	4938	62
Smote	522	522	4938	62

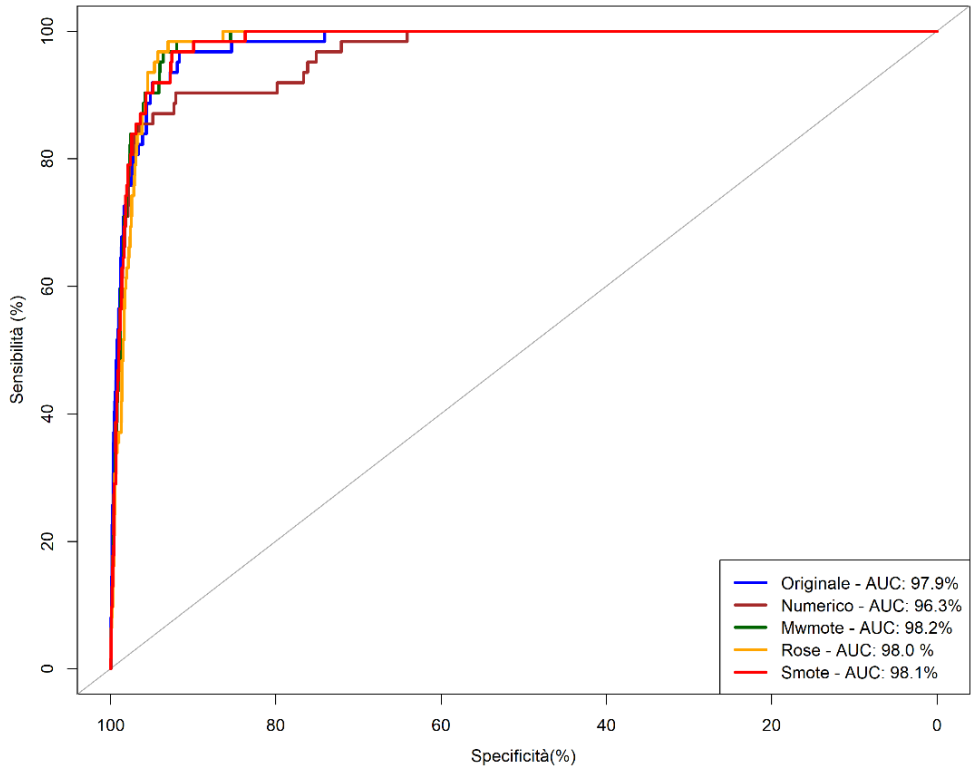
Al fine di valutare sia l'effetto di tale esclusione che l'effetto del bilanciamento, i modelli sono stati costruiti sul training set completo (*Originale*), sul training set non bilanciato e costituito dalle variabili quantitative (*Numerico*) e sui tre training set bilanciati.

4.4.2 Regressione Logistica

Bilanciamenti	Sensibilità	Specificità	Precisione	F1	Accuratezza
Originale	20,97%	99,82%	59,09%	30,95%	98,84%
Numerico	16,13%	99,76%	45,45%	23,81%	98,72%
Mwmote	90,32%	94,86%	18,06%	30,11%	94,80%
Rose	96,77%	93,07%	14,93%	25,86%	93,12%
Smote	91,94%	94,39%	17,07%	28,79%	94,36%

Per quanto concerne la regressione logistica, l'esclusione delle variabili qualitative ha generato un effetto negativo su qualsiasi indicatore, in particolare sulla precisione del modello che è diminuita del 13,64%. In termini di accuratezza e specificità entrambi i modelli non bilanciati hanno raggiunto performance piuttosto elevate. In questo caso, il valore dell'accuratezza è fuorviante per la valutazione del modello, poiché sarebbe possibile raggiungere lo stesso risultato anche nel caso in cui tutte le osservazioni del test set fossero classificate come "non acquisto". Inoltre, lo sbilanciamento ha causato performance piuttosto negative in termini di sensibilità e F1. Le tecniche di bilanciamento hanno permesso di aumentare notevolmente la sensibilità, raggiungendo il 96,77% con l'utilizzo del metodo Rose, e di mantenere performance elevate in termini di specificità e accuratezza. Tali indicatori sono diminuiti solamente del 4/5%, mentre la precisione è calata notevolmente. Si tratta di un risultato atteso poiché aumentando i veri positivi per la classe di minoranza, è aumentato anche il numero di falsi positivi[20]. Per tale ragione è possibile osservare il valore di F1, che combina sensibilità e precisione. A parità di variabili il bilanciamento ha comportato un aumento di F1, nonostante le performance siano inferiori a quelle osservate nel modello *Originale*. Nel confronto tra i metodi di bilanciamento si osserva che *Rose* ha raggiunto le performance più elevate in termini di sensibilità, mentre *Mwmote* ha permesso di ottenere risultati migliori per gli altri indicatori.

[20] Chawla, N. V., (2005) Data mining for imbalanced datasets: an overview.

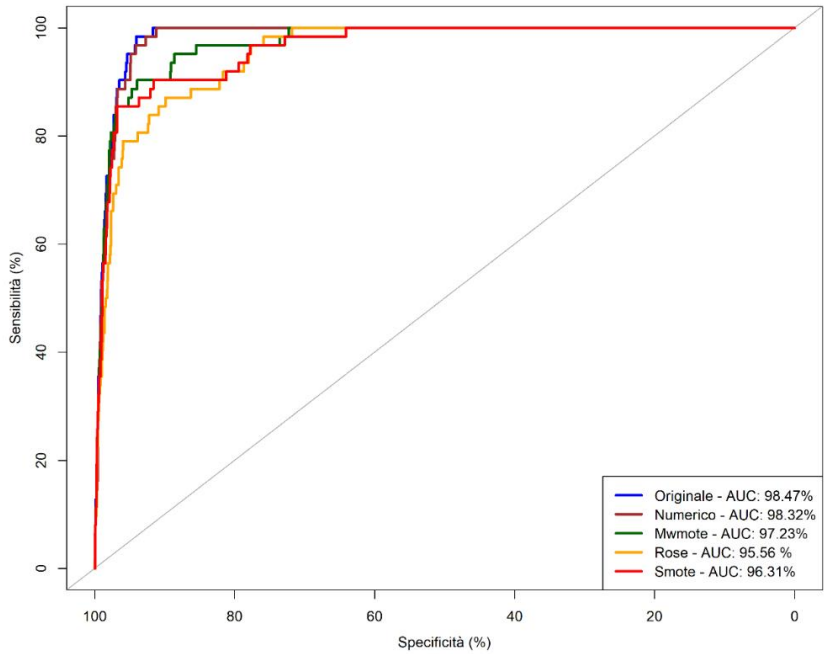


Dalle curve ROC si osserva che il modello *Numerico* consegue le performance peggiori anche relativamente all'AUC. Gli altri modelli hanno raggiunto prestazioni simili e il bilanciamento ha permesso di ottenere i valori dell'AUC più elevati, nonostante l'esclusione delle variabili qualitative abbia inizialmente comportato una diminuzione del 1,6%. I risultati migliori si ottengono con un valore soglia pari a 0.26 nel modello *Mwmote,* conseguendo una sensibilità del 98,39% e una specificità del 92,02%.

4.4.3 Analisi discriminante lineare

Bilanciamenti	Sensibilità	Specificità	Precisione	F1	Accuratezza
Originale	64,52%	98,54%	35,71%	45,98%	98,12%
Numerico	58,06%	98,44%	31,86%	41,14%	97,94%
Mwmote	87,10%	94,96%	17,82%	29,59%	94,86%
Rose	79,03%	94,43%	15,12%	25,39%	94,24%
Smote	85,48%	94,55%	16,46%	27,60%	94,44%

L'analisi discriminante lineare conferma che l'esclusione delle variabili qualitative produce un effetto negativo sulle prestazioni, anche se non si osservano variazioni consistenti.
Il bilanciamento comporta un aumento della sensibilità e contemporaneamente determina la diminuzione degli altri valori osservati, raggiungendo le migliori performance nel modello *Mwmote.* Dal confronto con i risultati ottenuti dalla regressione logistica si osserva che i modelli non bilanciati hanno conseguito performance più elevate in termini di sensibilità e F1, mentre gli altri valori sono diminuiti. Inoltre, il miglioramento della sensibilità generato dai metodi di bilanciamento è inferiore, sebbene gli altri valori siano pressoché analoghi.

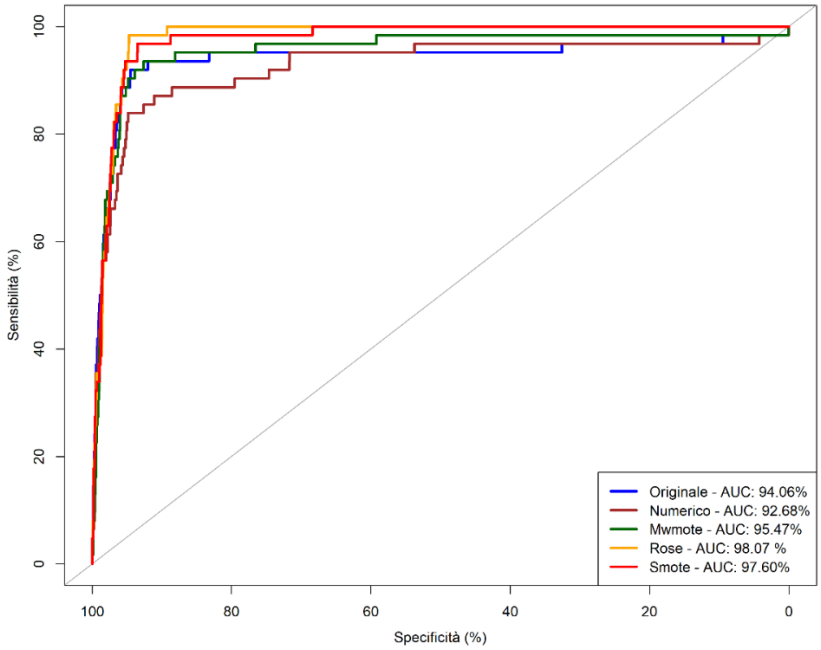


Le tecniche di bilanciamento hanno causato una diminuzione dell'AUC, fino al 95.56% nel modello *Rose.* Nei modelli non bilanciati le performance sono piuttosto elevate e l'AUC è superiore ai valori osservati nei modelli logit analizzati in precedenza. L'utilizzo delle variabili qualitative ha un effetto positivo sull'AUC, nonostante la differenza tra il modello *Originale* e quello *Numerico* sia inferiore all'1%.

4.4.4 Support Vector Machines

Bilanciamenti	Sensibilità	Specificità	Precisione	F1	Accuratezza
Originale	3,22%	99,94%	40,00%	5,97%	98,74%
Numerico	1,61%	99,98%	50,00%	3,13%	98,76%
Mwmote	87,10%	95,22%	18,62%	30,68%	95,12%
Rose	98,39%	92,97%	14,95%	25,96%	93,04%
Smote	96,77%	92,65%	14,18%	24,74%	92,70%

Attraverso i risultati delle support vector machines si evidenzia nuovamente che l'accuratezza è fuorviante per la valutazione del modello. Infatti, nonostante il modello *Originale* e *Numerico* siano particolarmente accurati, non sono in grado di prevedere i valori positivi: la sensibilità è pari al 3,22% e al 1,61%. Le tecniche di bilanciamento producono un aumento considerevole della sensibilità, fino a raggiungere il 98,39% nel modello *Rose,* e la riduzione della precisione, confermando che tali valori variano in maniera inversamente proporzionale. Dal valore di F1 si osserva un miglioramento delle prestazioni per effetto del bilanciamento e una riduzione a seguito dell'esclusione delle variabili qualitative.



I valori dell'AUC evidenziano un miglioramento delle performance con l'utilizzo delle tecniche di bilanciamento. L'esclusione delle variabili qualitative produce lo stesso effetto osservato nella regressione logistica, cioè una diminuzione dell'AUC inferiore al 2%. In questo caso le performance migliori si ottengono con il metodo *Rose,* conseguendo un valore dell'AUC pari al 98,07%.

Conclusioni

Nel corso di questo elaborato è stata svolta un'analisi con l'obiettivo di mostrare le differenze in termini di prestazioni di tre modelli di apprendimento automatico: regressione logistica, analisi discriminante lineare e support vector machines. Inoltre, è stata rivolta particolare attenzione alle tecniche di bilanciamento e al loro effetto sulle performance dei classificatori. Tali tecniche hanno ricevuto una notevole attenzione per la risoluzione delle problematiche legate ai dataset con classi di risposta sbilanciate, perciò è stava valutata la loro efficacia nel dataset del Google Store, in cui la classe di minoranza rappresenta solamente l'1,6% delle osservazioni. Nonostante ciascun classificatore abbia ottenuto un'accuratezza pari a circa il 98%, allo stesso tempo nessuno di questi è riuscito a classificare correttamente la classe degli acquisti. In particolare, le support vector machines hanno raggiunto il risultato più basso in termini di sensibilità, pari al 1,61% nel dataset numerico e al 3,22% nel dataset che considera tutte le variabili. Il modello che ha permesso di ottenere le migliori prestazioni in termini di sensibilità, nonostante lo sbilanciamento delle classi di risposta, è l'analisi discriminante lineare, con un valore pari al 58,06% nel dataset numerico e al 64,52% nel dataset completo. L'utilizzo delle tecniche di bilanciamento e di indicatori di performance alternativi all'accuratezza hanno permesso di evidenziare i seguenti aspetti:

- il numero di casi positivi classificati correttamente è aumentato in maniera esponenziale, fino a raggiungere una variazione del 97% con la tecnica MWMOTE per le support vector machines. A parità di variabili, il valore di F1 è migliorato per la regressione logistica e per le support vector machines;

- contemporaneamente la precisione si è ridotta drasticamente, a causa del maggior numero di falsi positivi, confermando quanto emerso dall'analisi della letteratura. L'accuratezza e la specificità si sono ridotte lievemente, mantenendo prestazioni piuttosto elevate;

- il valore dell'AUC mostrato dalle curve ROC è aumentato, con un'eccezione per l'analisi discriminante lineare in cui è stato raggiunto il miglior risultato prima del bilanciamento del dataset.

Le tecniche di bilanciamento hanno permesso di migliorare notevolmente la sensibilità dei modelli di classificazione ed è stato confermato che l'accuratezza è un indicatore fuorviante per la valutazione delle performance nei dataset sbilanciati, perciò è necessario affidarsi a misure alternative che consentono di ottenere una maggiore affidabilità.

Bibliografia

Barua S., Islam M. M, Yao X. & Murase K., (2014), MWMOTE- Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Transactions on Knowledge and Data Engineering, 26*, 405-425.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*, 1145–1159.

Chawla, N. V., Bowyer K. W., Hall L. O. & Kegelmeyer W. P., (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16,* 321-357

Chawla, N. V. (2005). Data mining for imbalanced datasets: an overview. In: Maimon O. & Rokach L. (Eds.). *Data Mining and Knowledge Discovery Handbook* (pp. 853-867). Springer.

Cordón I., García S., Fernández A. & Herrera F. (2018). Imbalance: Oversampling algorithms for imbalanced classification in R. *Knowledge-Based Systems, 161*, 329-341

Grolemund G. & Wickham H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software, 40(3),* 1–25.

Grosjean P. & Ibanez F. (2014) *Pastecs: Package for Analysis of Space-Time Ecological Series.*

Hand D.J. & Vinciotti V. (2003). Choosing K for Two-Class Nearest Neighbour Classifiers with Unbalanced Classes. *Pattern Recognition Letters, 24,* 1555-1562

He H. & Garcia E. A., (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21,* 1263-1284

Honaker J., King G. & Blackwell M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software, 45(7),* 1–47.

J L. (2006). Plotrix: a package in the red light district of R. *R-News, 6(4),* 8-12.

James G., Witten D., Hastie T. & Tibshirani R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

King G. & Zeng L., (2001). Logistic regression in rare events data. *Political Analysis, 9(2),* 137-163.

Kotsiantis S., Kanellopoulos D. & Pintelas P., (2006). Handling imbalanced dataset: A review. *GESTS International Transactions on Computer Science and Engineering*, *30,* 25-36.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, 28(5),* 1-26.

Lunardon N., Menardi G. & Torelli N. (2014). ROSE: a Package for Binary Imbalanced Learning*. R Journal, 6(1)*, 82-92.

Mauriello F., (2014), *Tecniche di ricampionamento per dataset con classi di risposta sbilanciate. Una proposta metodologica per dataset con predittori di natura numerica e categorica.*

McKinsey Global Institute (2016), *The Age Of Analytics: Competing In A Data-Driven World.*

Menardi G. & Torelli N. (2012). Training and assessing classification rules with unbalanced data. *Data Mining and Knowledge Discover.*

Meyer D. et al. (2020). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071).* TU Wien.

Neuwirth E. (2014). *RColorBrewer: ColorBrewer Palettes.*

Ooms J. (2014). *The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects.*

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL https://www.R-project.org/.

Swets J., (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, *240(4857),* 1285–1293.

Torgo L. (2010). *Data Mining with R, learning with case studies Chapman and Hall/CRC*.

Tuszynski J. (2020). *caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc*.

Venables, W. N. & Ripley, B. D. (2002*) Modern Applied Statistics with S. Fourth Edition.* Springer.

Wickham H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag

Wickham H. et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4(43),* 1686

Wickham H., François R., Henry L. & Müller K. (2021). *dplyr: A Grammar of Data Manipulation*.

Wu G. & Chang E., (2003). Class-Boundary Alignment for Imbalanced Dataset Learning. *ICML 2003 Workshop on Learning from Imbalanced Data Sets*.

Robin X. et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77.